Journal of Multivariate Analysis 101 (2010) 1574–1593

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Statistical inference of minimum BD estimators and classifiers for varying-dimensional models

Chunming Zhang

Department of Statistics, University of Wisconsin, 53706 Madison, WI, United States

ARTICLE INFO

Article history:
Received 26 February 2009
Available online 4 March 2010

AMS 2000 subject classifications:
primary 62F12
62F05
secondary 62J02
60J12

Keywords:
A diverging number of parameters
Exponential family
Hemodynamic response function
Loss function
Optimal Bayes rule

ABSTRACT

Stochastic modeling for large-scale datasets usually involves a varying-dimensional model space. This paper investigates the asymptotic properties, when the number of parameters grows with the available sample size, of the minimum-BD estimators and classifiers under a broad and important class of Bregman divergence (BD), which encompasses nearly all of the commonly used loss functions in the regression analysis, classification procedures and machine learning literature. Unlike the maximum likelihood estimators which require the joint likelihood of observations, the minimum-BD estimators are useful for a range of models where the joint likelihood is unavailable or incomplete. Statistical inference tools developed for the class of large dimensional minimum-BD estimators and related classifiers are evaluated via simulation studies, and are illustrated by analysis of a real dataset.

1. Introduction

In many fields of applications, the dimension (number) p of model space (parameters) depends on the sample size n. Examples include the X-ray crystallography [1,2] and the autoregressive models in times series. In the literature, Drost [3] developed the goodness-of-fit tests for location-scale models when the number of classes tends to infinity; Murphy [4] developed testing for a time dependent coefficient in Cox's regression model. This paper is motivated from issues in two important and challenging applications.

1.1. fMRI time series: a diverging number of parameters

Functional magnetic resonance imaging (fMRI) is a recent and exciting method that allows investigators to determine which areas of the brain are involved in a cognitive task. Following Ward [5] and Worsley et al. [6], a single-voxel fMRI time-series $\{s(t_i), y(t_i)\}_{i=1}^n$, for a given scan and a given subject, can be captured by the convolution model

$$y(t) = d(t) + s * h(t) + \varepsilon(t), \quad t = t_1, \ldots, t_n, \tag{1.1}$$

where $*$ denotes the convolution operator, $y(t)$ is the measured noisy fMRI signal, $s(t)$ is the external input stimulus (which could be from a design either block- or event-related and where $s(t) = 1$ or 0 indicates the presence or absence of a stimulus), $h(t)$ is the hemodynamic response function (HRF) at time t after neural activity, $d(t)$ is a slowly drifting baseline, and the errors $\varepsilon(t_i)$ are zero-mean and temporally autocorrelated. Similar models can be found in [7]. Refer to [8] and references therein for a recent review of statistical issues and methods in fMRI data analysis.

E-mail address: cmzhang@stat.wisc.edu.

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2010.02.009

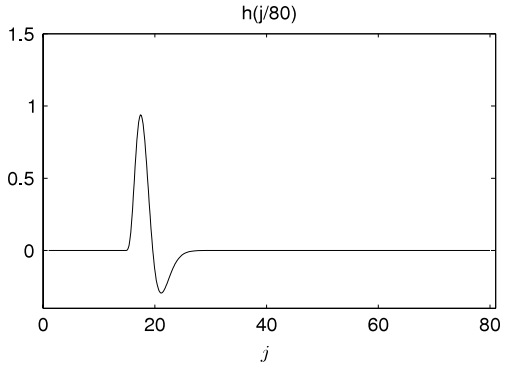

Fig. 1. An illustrative plot of HRF $h(t_j)$ with $n = 80$.

Of primary interest to neuroscientists is the estimation and hypothesis testing of the underlying HRF. Typically, the peak value of the HRF $h(\cdot)$ is reached after a short delay of the stimulus and drops quickly to zero. A typical example of $h(\cdot)$, given in [9], is plotted in Fig. 1. Clearly, the region $\{t : h(t) \neq 0\}$ is sparse in its temporal domain. Thus, to obtain statistically more efficient estimates of the HRF associated with event-related fMRI experiments, the sparsity of the HRF needs to be taken into account. We thus suppose that $h(t) = 0$ for $t > t_{p_n}$ and focus on estimating the first p_n values of $h(t_i)$, where p_n is less than n, the length of the fMRI time series. In neuroimaging studies, the temporal drift $d(\cdot)$ is a nuisance function and usually approximated by a (at most third order) polynomial; see for example, the popular imaging analysis tool AFNI at http://afni.nimh.nih.gov/afni/ [10,6]. As such, (1.1) is re-expressed as

$$\mathbf{y} = \widetilde{\mathbf{T}}\widetilde{\boldsymbol{\alpha}} + \mathbf{S}\mathbf{h} + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\mathbf{y} = (y(t_1), \ldots, y(t_n))^T$,

$$\widetilde{\mathbf{T}} = \begin{bmatrix} 1 & t_1 & t_1^2 & t_1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 & t_n^3 \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} s(0) & 0 & \cdots & 0 \\ s(t_2 - t_1) & s(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ s(t_{p_n} - t_1) & s(t_{p_n} - t_2) & \cdots & s(0) \\ \vdots & \vdots & \cdots & \vdots \\ s(t_n - t_1) & s(t_n - t_2) & \cdots & s(t_n - t_{p_n}) \end{bmatrix}$$

is the $n \times p_n$ Toeplitz matrix, $\boldsymbol{\epsilon} = (\epsilon(t_1), \ldots, \epsilon(t_n))^T$, and $\widetilde{\boldsymbol{\alpha}} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^T$ and $\mathbf{h} = (h(t_1), \ldots, h(t_{p_n}))^T$ are both vectors of unknown parameters.

Clearly, p_n here grows with n but is excessively below n. For the regression problem (1.2), the large-dimensional vector $\mathbf{h}$ can be estimated via weighted least-squares using the quadratic loss. In practice, however, p_n, termed the "intrinsic dimensionality of fMRI data" [11], is unknown for real fMRI data. Indeed, as far as we know, all published work for fMRI assumes that $p_n = p$ is a known fixed number, followed by traditional parametric inference based on asymptotic derivations of fixed-dimensional estimators. To reduce modeling biases due partly to the fixed choice of p, statistical inference based on asymptotic results which allow the dimension p_n to depend on n is desired. This motivates us to consider a more realistic relation between p_n and n,

$$\text{dimension} : p_n \text{ varies with } n \text{ or even } p_n \to \infty \text{ at a certain rate as } n \to \infty. \tag{1.3}$$

1.2. Statistical learning: regression and classification under general loss

In statistical learning, the primary goals of regression and classification seem to be kept separate. Regression methods concern the "orderable" output variable and aim to estimate the regression function at points of the input variable, whereas the primary interest of classification rules for the "categorical" output variable is to forecast the most likely class label for the output.

As discussed in [12], both regression and classification can be viewed from the common perspective of real valued prediction. Namely, the goal of a supervised learning algorithm is to use the training samples to construct a prediction rule for a future output at the observed value of the input variable. Depending on the nature of the output variable, the predictive error is quantified by different error measures. For example, the quadratic loss function, as utilized in the previous brain fMRI data, has nice analytical properties and is usually used in regression analysis. However, the quadratic loss is not always adequate in classification problems where the misclassification loss, deviance loss (or the negative log-likelihood) and exponential loss are more realistic and commonly used in classification.

Owing to the nature of output variables in classification, the choice of loss functions plays an important role in defining and understanding the bias, variance and prediction error for the classification rule [13]. Recent work in the area of research includes James [14], Efron [15], Bartlett, et al. [16], and many others. This inspires us to consider a more general framework,

$$\text{loss}: \text{ the loss function belongs to the class of Brègman divergence (BD)}, \tag{1.4}$$

where the notion BD, as exemplified in Section 2.1, unifies nearly all of the commonly used loss functions in the regression analysis, classification procedures and machine learning literature.

For fixed dimensions, the asymptotic properties (including the consistency and asymptotic distribution) of minimum-BD estimators can be derived using the existing results for M-estimators. However, for diverging dimensions, asymptotic properties of the M-estimators based on the empirical process theory may not be fully transparent. This paper aims at obtaining the asymptotics of minimum-BD estimators with the diverging dimensions, in an accessible way.

1.3. Outline of this paper

In this paper, we aim to investigate the unified statistical properties and develop the powerful inference tools relevant for the class of large-dimensional minimum-BD estimators under the Brègman divergence framework. By doing this task, we hope to provide new insights into the statistical behaviors of both regression estimators and classification procedures for all fields of scientific research that concern large-dimensional problems.

This paper differs from existing work in a number of ways. First, this paper integrates the important loss function in (1.4) with dimensions in (1.3) simultaneously. Portnoy [17] considered (1.3) for the maximum-likelihood estimator, from an i.i.d. sample $\{\mathrm{X}_{n1}, \ldots, \mathrm{X}_{nn}\}$, having a distribution in the exponential family, i.e. the loss function is the negative log-likelihood. Fan and Peng [18] considered (1.3) for the penalized likelihood estimator, from an i.i.d. sample $\{(\mathrm{X}_{n1}, Y_{n1}), \ldots, (\mathrm{X}_{nn}, Y_{nn})\}$, where the loss function is the negative log-likelihood. Second, as noted, likelihood-based results are not directly applicable in the context that the standard likelihood is either unavailable or difficult to optimize. Accordingly, the above existing work is not applicable to statistical inference for large-dimensional maximum "quasi-likelihood" estimation, for example, but our study can immediately be useful. See Section 2.1. Third, from the classification viewpoint, our study elucidates the applicability and consistency of many non-likelihood based minimum-BD classifiers, despite the lack of efficiency of the corresponding minimum-BD regression estimators. Thus, one is able to readily assess the impact of loss functions on the performance of various classifiers induced by either likelihood or non-likelihood based estimators, when applied to large-dimensional datasets. In contrast, results confined to likelihood-based estimation cannot achieve this goal. Fourth, the minimum-BD estimator is allowed to be a local minimizer and its rate of consistency and limit distribution can be obtained, when the dimension grows to infinity; see Theorem 1. This differs from the well-known result for the M-estimator [19–21], where the dimension is fixed for the local minimizer, rate of consistency, and asymptotic distribution.

What is special about the minimum-BD estimators as opposed to M-estimators? The BD is partly motivated from the machine learning literature, where most of the loss functions that people are specifically interested in belong to the family of BD, and offers many practical advantages. For example, constructing the loss function for multi-class classification is challenging. On the other hand, if we know the generating q-function for binary-classification, then it can conveniently be generalized to the counterpart for multi-class classification. Accordingly, the corresponding Q-loss for multi-class classification can be obtained. That is, the BD focuses on the aspect of constructing loss functions, whereas an M-estimation procedure is to estimate parameters assuming that the loss is available.

The rest of the paper is arranged as follows. Section 2 introduces the BD and formulates the minimum-BD estimator. Section 3 investigates the statistical sampling properties of the minimum-BD estimator. Section 4 conducts hypothesis testing for the model parameters based on the minimum-BD estimator. For binary classification, Section 5 establishes the consistency of the minimum-BD classifier. Section 6 presents simulation studies, and Section 7 applies results developed in Sections 3–5 to real data. Section 8 ends the paper with a brief discussion. All technical details are relegated to the Appendix.

2. Minimum-BD estimator

2.1. Brègman divergence

Brègman [22] introduced a device for constructing a bivariate function,

$$Q(v, \mu) = -q(v) + q(\mu) + (v - \mu)q^{(1)}(\mu),$$

for a given concave q-function. Note that the concavity requirement on q ensures the non-negativity of Q. However, since $Q(v, \mu)$ is not generally symmetric in arguments, Q is not a "metric" or "distance" in the strict sense. Hence, we call Q the "Brègman divergence" (BD) and call q the "generating function" of Q.

It is easy to see that, with the flexible choice of the q-function, the BD is suitable for a broad class of error measures. Below we present some notable examples of the Q-loss constructed from the q-function. A function $q(\mu) = a\mu - \mu^2$ for some constant a yields the quadratic loss $Q(Y, \mu) = (Y - \mu)^2$. For a binary response variable Y, $q(\mu) = \min\{\mu, (1 - \mu)\}$ gives the misclassification loss $Q(Y, \mu) = \mathrm{I}\{Y \neq \mathrm{I}[\mu > 1/2]\}$; $q(\mu) = -\{\mu \log(\mu) + (1 - \mu)\log(1 - \mu)\}$ gives the

Bernoulli deviance loss $Q(Y,\mu) = -\{Y\log(\mu) + (1-Y)\log(1-\mu)\}$; $q(\mu) = 2\min\{\mu, (1-\mu)\}$ results in the hinge loss $Q(Y,\mu) = \max\{1-(2Y-1)\text{sign}(\mu-0.5), 0\}$ of the support vector machine; $q(\mu) = 2\{\mu(1-\mu)\}^{1/2}$ yields the exponential loss $Q(Y,\mu) = \exp[-(Y-0.5)\log\{\mu/(1-\mu)\}]$ used in [23]. Among many others, the quasi-likelihood function [24], and the Kullback–Leibler divergence (or the deviance loss) for the exponential family of probability functions fall into the class of BD.

2.2. Statistical model and the minimum-BD estimator

Let $\{(\mathtt{X}_{ni}, Y_{ni})\}_{i=1}^n$ be a sample of i.i.d. observations from the population $(\mathtt{X}_n, Y_n)$, where $\mathtt{X}_{ni} = (X_{i1}, \ldots, X_{ip_n})^T$ is the p_n-dimensional input vector corresponding to the ith output variable Y_{ni} and p_n is known. Suppose that the mean regression function, $m(\mathtt{x}_n) = E(Y_n \mid \mathtt{X}_n = \mathtt{x}_n)$, is described by the model,

$$m(\mathtt{x}_n) = F^{-1}(\beta_{n,0;0} + \mathtt{x}_n^T\boldsymbol{\beta}_{n;0}), \tag{2.1}$$

where $\beta_{n,0;0} \in \mathbb{R}^1$ and $\boldsymbol{\beta}_{n;0} = (\beta_{n,1;0}, \ldots, \beta_{n,p_n;0})^T \in \mathbb{R}^{p_n}$ stand for the unknown "true" model parameters, and $F(m)$ is a known link function, oftentimes bearing a monotonic relation with m.

Our goal is to estimate the true parameters $(\beta_{n,0;0}, \boldsymbol{\beta}_{n;0})$. Under the general loss Q, the minimum-BD estimator $(\widehat{\beta}_{n,0}, \widehat{\boldsymbol{\beta}}_n)$ is defined as the minimizer of the criterion function,

$$\ell_n(\beta_{n,0}, \boldsymbol{\beta}_n) = \frac{1}{n}\sum_{i=1}^n Q(Y_{ni}, F^{-1}(\beta_{n,0} + \mathtt{X}_{ni}^T\boldsymbol{\beta}_n)),$$

where $\boldsymbol{\beta}_n = (\beta_{n,1}, \ldots, \beta_{n,p_n})^T$ and the loss $Q(\cdot,\cdot)$ belongs to the class BD. Set $\widetilde{\boldsymbol{\beta}}_n = (\beta_{n,0}, \boldsymbol{\beta}_n^T)^T$ and correspondingly $\widetilde{\mathtt{X}}_{ni} = (1, \mathtt{X}_{ni}^T)^T$. Then the criterion function above can be written as

$$\ell_n(\widetilde{\boldsymbol{\beta}}_n) = \frac{1}{n}\sum_{i=1}^n Q(Y_{ni}, F^{-1}(\widetilde{\mathtt{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n)). \tag{2.2}$$

The minimum-BD estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_n = (\widehat{\beta}_{n,0}, \widehat{\beta}_{n,1}, \ldots, \widehat{\beta}_{n,p_n})^T$ amounts to $\widehat{\widetilde{\boldsymbol{\beta}}}_n = \arg\min_{\widetilde{\boldsymbol{\beta}}_n} \ell_n(\widetilde{\boldsymbol{\beta}}_n)$.

We wish to emphasize here that the construction of a minimum-BD estimator does not necessarily need the distributional assumption about $(\mathtt{X}_n, Y_n)$. Therefore, when the likelihood function is unknown or is unduly complicated to optimize, acquiring the minimum-BD estimator seems more natural and viable.

3. Properties of the minimum-BD estimator

This section begins by studying the consistency and asymptotic distribution of the minimum-BD estimator. Unless otherwise stated, $\|\cdot\|$ denotes the L_2-norm.

3.1. Consistency

Define by $\widetilde{\boldsymbol{\beta}}_{n;0} = (\beta_{n,0;0}, \boldsymbol{\beta}_{n;0}^T)^T$ the vector of true regression parameters. Theorem 1 guarantees the existence of a consistent local minimizer for (2.2).

Theorem 1 (*Existence and Consistency*). *Assume Condition* A *in the Appendix. If* $p_n^4/n \to 0$ *as* $n \to \infty$, *then there exists a local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ *of* $\ell_n(\widetilde{\boldsymbol{\beta}}_n)$ *such that* $\|\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}\| = O_P(\sqrt{p_n/n})$.

Theorem 1 indicates that the local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ is $\sqrt{n/p_n}$-consistent. Assume that

$$\mathtt{q}_j(y;\theta) = (\partial^j/\partial\theta^j)Q(y, F^{-1}(\theta)), \quad j = 0, 1, \ldots \tag{3.1}$$

exist finitely up to any order required. Regarding the uniqueness of $\widehat{\widetilde{\boldsymbol{\beta}}}_n$, provided that

$$\mathtt{q}_2(y;\theta) > 0 \quad \text{for all } \theta \in \mathbb{R} \text{ and all } y \text{ in the range of the response variable,} \tag{3.2}$$

the criterion function $\ell_n(\widetilde{\boldsymbol{\beta}}_n)$ is convex in $\widetilde{\boldsymbol{\beta}}_n$, and hence the local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ is also the unique global minimum-BD estimator.

3.2. Asymptotic normality

Following Theorem 1, the asymptotic normality of the local minimizer is given in Theorem 2 below. Before stating it, we first introduce some necessary notations. Let $\widetilde{\mathrm{X}}_n = (1, \mathrm{X}_n^T)^T$,

$$\Omega_n = E\left[\mathrm{var}(Y_n \mid \mathrm{X}_n)\frac{\{q^{(2)}(m(\mathrm{X}_n))\}^2}{\{F^{(1)}(m(\mathrm{X}_n))\}^2}\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T\right], \qquad \mathbf{H}_n = -E\left[\frac{q^{(2)}(m(\mathrm{X}_n))}{\{F^{(1)}(m(\mathrm{X}_n))\}^2}\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T\right].$$

Theorem 2 (*Asymptotic Normality*). *Assume Condition* B *in the Appendix. If* $p_n^5/n \to 0$ *as* $n \to \infty$, *then any* $\sqrt{n/p_n}$*-consistent local minimizer* $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ *satisfies: for any fixed integer* $\mathrm{k} \geq 1$ *and any* $\mathrm{k} \times (p_n + 1)$ *matrix* A_n *such that* $A_nA_n^T \to G$ *with* G *being a* $\mathrm{k} \times \mathrm{k}$ *nonnegative-definite symmetric matrix,* $\sqrt{n}A_n\Omega_n^{-1/2}\mathbf{H}_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, G)$.

Theorem 2 reveals that the asymptotic distribution of the p_n-dimensional parametric estimator under BD depends on the choice of the loss function Q only through the second derivative of its generating q-function.

3.2.1. Lower bound of the asymptotic covariance matrices

According to Theorem 2, the asymptotic covariance matrix of $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ is $V_n = \mathbf{H}_n^{-1}\Omega_n\mathbf{H}_n^{-1}$. Is there an optimal choice of the q-function such that V_n achieves its lower bound? Proposition 1 manifests that the optimal q-function satisfies

$$q^{(2)}(m(\cdot)) = -\frac{c}{\mathrm{var}(Y_n \mid \mathrm{X}_n = \cdot)}, \quad \text{for a constant } c > 0. \tag{3.3}$$

Proposition 1. *If the q-function satisfies* (3.3), *then* V_n *achieves the lower bound*

$$\left(E[1/\mathrm{var}(Y_n \mid \mathrm{X}_n)\{F'(m(\mathrm{X}_n))\}^{-2}\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T]\right)^{-1}.$$

Remark 1. Under a BD Q, condition (3.3) in the case of $c = 1$ is equivalent to

$$E\left\{\frac{\partial^2 Q(Y_n, m(\cdot))}{\partial m(\cdot)^2} \mid \mathrm{X}_n = \cdot\right\} = E\left[\left\{\frac{\partial Q(Y_n, m(\cdot))}{\partial m(\cdot)}\right\}^2 \mid \mathrm{X}_n = \cdot\right],$$

which includes the conventional Bartlett identity [25] as a special case, when Q is the negative log-likelihood. Thus, we call (3.3) the "*generalized Bartlett identity*". It is also seen that the quadratic loss satisfies (3.3) for homoscedastic regression models even without knowing the error distribution.

3.2.2. Consistent asymptotic covariance matrix estimation

In many real applications, V_n is unknown. To conduct statistical inference for the true parameters $\widetilde{\boldsymbol{\beta}}_{n;0}$, V_n needs to be estimated. See the next section for the use of an estimated V_n in the proposed generalized Wald type test statistic. Typically, the sandwich formula can be exploited to form an estimator of V_n by

$$\widehat{V}_n = \widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1},$$

where

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{i=1}^n \mathrm{q}_1^2(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}}_n)\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T, \qquad \widehat{\mathbf{H}}_n = \frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}}_n)\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T.$$

Remark 2. In the particular case of homoscedastic regression models, applying the quadratic loss and the identity link is natural, and hence, the direct use of $4\widehat{\sigma}^2(n^{-1}\sum_{i=1}^n \widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T)$, where $\widehat{\sigma}^2 = \sum_{i=1}^n (Y_{ni} - \widetilde{\mathrm{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}}_n)^2/\{n - (p_n + 1)\}$, is more efficient than the above $\widehat{\Omega}_n$.

Proposition 2 below demonstrates that for any $\sqrt{n/p_n}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ of $\widetilde{\boldsymbol{\beta}}_{n;0}$, $\widehat{V}_n$ is a consistent estimator for V_n, in the sense that $A_n(\widehat{V}_n - V_n)A_n^T \xrightarrow{P} \mathbf{0}$ for any $\mathrm{k} \times (p_n + 1)$ matrix A_n satisfying $A_nA_n^T \to G$ (where k is any fixed integer). In the special case that the dimension p_n does not depend on n, the consistency of $\widehat{V}_n$ to V_n corresponds to $\widehat{V}_n - V_n \xrightarrow{P} \mathbf{0}$, the conventional notion of consistency.

Proposition 2 (*Covariance Matrix Estimation*). *Assume Condition* B *in the Appendix. If* $p_n^4/n \to 0$ *as* $n \to \infty$, *then for any* $\sqrt{n/p_n}$*-consistent estimator* $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ *of* $\widetilde{\boldsymbol{\beta}}_{n;0}$, *we have* $A_n(\widehat{V}_n - V_n)A_n^T \xrightarrow{P} \mathbf{0}$ *for any* $\mathrm{k} \times (p_n + 1)$ *matrix* A_n *satisfying* $A_nA_n^T \to G$, *where* G *is a* $\mathrm{k} \times \mathrm{k}$ *matrix and* k *is any fixed integer.*

4. Hypothesis testing via the minimum-BD estimator

This section conducts the hypothesis testing of the true parameters $\widetilde{\boldsymbol{\beta}}_{n;0}$, and derives the asymptotic distributions of the proposed test statistic.

4.1. Under the null hypothesis

We consider the hypothesis testing about $\widetilde{\boldsymbol{\beta}}_{n;0}$ formulated as

$$H_0 : A_n\widetilde{\boldsymbol{\beta}}_{n;0} = \mathbf{0} \quad \text{versus} \quad H_1 : A_n\widetilde{\boldsymbol{\beta}}_{n;0} \neq \mathbf{0}, \tag{4.1}$$

where A_n is a given $\mathsf{k} \times (p_n + 1)$ matrix such that $A_nA_n^T = G$ with G being a $\mathsf{k} \times \mathsf{k}$ positive-definite matrix. This form of linear hypotheses allows one to simultaneously test whether a few variables are statistically significant by taking some specific form of the matrix A_n, for example

$$A_n = [\mathbf{I}_{\mathsf{k}}, \mathbf{0}_{\mathsf{k},p_n+1-\mathsf{k}}] \tag{4.2}$$

yields $A_nA_n^T = \mathbf{I}_{\mathsf{k}}$.

We propose a generalized Wald type test statistic,

$$W_n = n(A_n\widehat{\widetilde{\boldsymbol{\beta}}}_n)^T(A_n\widehat{V}_nA_n^T)^{-1}(A_n\widehat{\widetilde{\boldsymbol{\beta}}}_n),$$

where $\widehat{V}_n$ is defined in the previous subsection. Theorem 3 justifies that under the null, W_n would for large n be distributed as χ^2_{k}.

Theorem 3 (*Wald Type Test Under H_0*). *Assume Condition* C *in the Appendix. If $p_n^5/n \to 0$ as $n \to \infty$, then under H_0 in (4.1), we have that $W_n \xrightarrow{\mathcal{L}} \chi^2_{\mathsf{k}}$ for any $\sqrt{n/p_n}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ of $\widetilde{\boldsymbol{\beta}}_{n;0}$.*

It is noted that the test statistic of the form,

$$\Lambda_n = 2n\left\{\min_{\widetilde{\boldsymbol{\beta}}_n \in \mathbb{R}^{p_n+1}: A_n\widetilde{\boldsymbol{\beta}}_n = \mathbf{0}} \ell_n(\widetilde{\boldsymbol{\beta}}_n) - \min_{\widetilde{\boldsymbol{\beta}}_n \in \mathbb{R}^{p_n+1}} \ell_n(\widetilde{\boldsymbol{\beta}}_n)\right\},$$

reduces to the classical likelihood-ratio test statistic, when the Q-loss in (2.2) is set to be the negative log-likelihood and the dimension p_n is fixed. In that case, it is well-known that Λ_n, in general, follows an asymptotic χ^2 distribution under the null. Theorem 4 below explores the extent to which this Wilks type of result can feasibly be extended to Λ_n constructed from the broad q-class of loss functions in the presence of a diverging number p_n of parameters.

Theorem 4 (*Likelihood-ratio Type Test Under H_0*). *Assume (3.2) and Condition* D *in the Appendix. If $p_n^5/n \to 0$ as $n \to \infty$ and the q-function satisfies (3.3), then under H_0 in (4.1), we have that $\Lambda_n/c \xrightarrow{\mathcal{L}} \chi^2_{\mathsf{k}}$ for any $\sqrt{n/p_n}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ of $\widetilde{\boldsymbol{\beta}}_{n;0}$.*

Curiously, the result in Theorem 4 indicates that the restrictive assumption (3.3) on the q-function limits the application domain of the test statistic Λ_n. For instance, in the particular case of binary responses, it is clearly seen that the Bernoulli deviance loss satisfies (3.3), but the quadratic loss and exponential loss violate (3.3). This limitation indeed reflects that under the general framework of BD, the likelihood-ratio type test statistic Λ_n may not be straightforwardly valid. Computationally, W_n is also easier to use than Λ_n. Thus, Section 4.2 below focuses on W_n.

4.2. Under the alternative hypothesis

To appreciate the discriminating power of W_n in assessing the significance, the asymptotic power is analyzed. Theorem 5 demonstrates that W_n is consistent against all fixed deviations from the null model.

Theorem 5 (*Wald Type Test Under H_1*). *Assume Condition* C *in the Appendix and $A_nV_nA_n^T \xrightarrow{P} \mathbf{M}$ where $\mathbf{M}$ is a $\mathsf{k} \times \mathsf{k}$ positive definite matrix. If $p_n^5/n \to 0$ as $n \to \infty$, then under the fixed alternative H_1 in (4.1) where $\|A_n\widetilde{\boldsymbol{\beta}}_{n;0}\|$ is independent of n, we have that*

$$n^{-1}W_n \geq \lambda^{-1}_{\max}(\mathbf{M})\|A_n\widetilde{\boldsymbol{\beta}}_{n;0}\|^2 + o_P(1)$$

for any $\sqrt{n/p_n}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ of $\widetilde{\boldsymbol{\beta}}_{n;0}$.

The result in Theorem 5 manifests that under the fixed alternative H_1, $W_n \xrightarrow{P} +\infty$ at the rate n. Hence W_n has the power function tending to one against fixed alternatives.

Consider a sequence of contiguous alternatives, given by

$$H_{1n} : A_n\widetilde{\boldsymbol{\beta}}_{n;0} = \delta_n \boldsymbol{c}\{1 + o(1)\}, \tag{4.3}$$

where $\delta_n = n^{-1/2}$ and $\boldsymbol{c} = (c_1, \ldots, c_{\mathtt{k}})^T \neq \mathbf{0}$ is fixed, namely, $\sqrt{n}A_n\widetilde{\boldsymbol{\beta}}_{n;0} \to \boldsymbol{c} \neq \mathbf{0}$ as $n \to \infty$. Theorem 6 explores the asymptotic distribution of W_n under the contiguous alternatives H_{1n}.

Theorem 6 (*Wald Type Test Under H_{1n}*). *Assume Condition* C *in the Appendix and $A_nV_nA_n^T \xrightarrow{P} \mathbf{M}$ where $\mathbf{M}$ is a $\mathtt{k} \times \mathtt{k}$ positive definite matrix. If $p_n^5/n \to 0$ as $n \to \infty$, then under the contiguous alternative H_{1n} in (4.3), we have that $W_n \xrightarrow{\mathcal{L}} \chi^2_{\mathtt{k}}(\tau^2)$ for any $\sqrt{n/p_n}$-consistent estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ of $\widetilde{\boldsymbol{\beta}}_{n;0}$, with the noncentrality parameter $\tau^2 = \boldsymbol{c}^T\mathbf{M}^{-1}\boldsymbol{c}$.*

The result in Theorem 6 suggests that W_n has a non-trivial local power detecting contiguous alternatives approaching the null at the rate $n^{-1/2}$.

5. Consistency of the minimum-BD classifier

This section deals with the binary response variable Y_n which only takes values either 0 or 1. In this case, the mean regression function $m(\mathtt{x}_n)$ in (2.1) becomes the class probability, $P(Y_n = 1 \mid \mathtt{X}_n = \mathtt{x}_n)$. From the minimum-BD estimator $(\widehat{\beta}_{n,0}, \widehat{\boldsymbol{\beta}}_n^T)^T$ proposed in Section 2, we can construct the following minimum-BD classifier,

$$\widehat{\phi}_n(\mathtt{x}_n) = \mathrm{I}\{\widehat{m}(\mathtt{x}_n) > 1/2\},$$

for a future input $\mathtt{x}_n$, where $\mathrm{I}(\cdot)$ is an indicator function and $\widehat{m}(\mathtt{x}_n) = F^{-1}(\widehat{\beta}_{n,0} + \mathtt{x}_n^T\widehat{\boldsymbol{\beta}}_n)$. Details on binary classification can be found in [26].

5.1. Consistency

To emphasize the dependence of the dimension p_n on n in our current setting, the optimal Bayes rule is denoted by $\phi_{n,\mathrm{B}}(\mathtt{x}_n) = \mathrm{I}\{m(\mathtt{x}_n) > 1/2\}$. For a test sample $(\mathtt{X}_n^o, Y_n^o)$, which is an i.i.d. copy of samples in the training set $\mathcal{T}_n = \{(\mathtt{X}_{ni}, Y_{ni}), i = 1, \ldots, n\}$, the optimal Bayes risk is then $R(\phi_{n,\mathrm{B}}) = P\{\phi_{n,\mathrm{B}}(\mathtt{X}_n^o) \neq Y_n^o\}$. Meanwhile, the conditional risk of the minimum-BD classification rule $\widehat{\phi}_n$ is $R(\widehat{\phi}_n) = P\{\widehat{\phi}_n(\mathtt{X}_n^o) \neq Y_n^o \mid \mathcal{T}_n\}$. For $\widehat{\phi}_n$ induced by the minimum-BD regression estimation using a range of loss functions, Theorem 7 verifies the classification consistency preserved by $\widehat{\phi}_n$.

Theorem 7 (*Consistency of the Minimum-*BD *Classifier*). *Assume Conditions* A1 *and* A4 *in the Appendix. Suppose that $\|\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}\| = O_P(r_n)$. If $r_n\sqrt{p_n} = o(1)$, then the classification rule $\widehat{\phi}_n$ constructed from $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ is consistent in the sense that $E\{R(\widehat{\phi}_n)\} - R(\phi_{n,\mathrm{B}}) \to 0$ as $n \to \infty$.*

Zhang [27] derived some non-asymptotic classification error bound, and obtained classification consistency under kernel formulations. There, both cases require the use of convex loss functions, though a parametric form of $P(Y_n = 1 \mid \mathtt{X}_n = \mathtt{x}_n)$ is not assumed. As a comparison, Theorem 7 does not require convexity of the loss; the parametric structure is used for illustration purposes only and can be relaxed.

6. Simulation studies

6.1. Inference of the HRF in fMRI over a single voxel

Typically, the analysis of the entire brain fMRI data is conducted by a two-step procedure: voxelwise statistical analysis in the first step, followed by multiple comparison in the second step. Since this paper aims at developing statistical inference tools in the first step, illustration of single-voxel analysis is focused on.

We simulate an fMRI experiment with a single run and a single type of stimulus. In the simulation, $n = 400$, $t_i = i/n$, $i = 1, \ldots, n$, and 1000 realizations are conducted. There are many different choices for p_n to be made in different contexts. For illustrative purpose only, we set $p_n = [9(n^{1/5.5} - 1)] + 1$ which, on one hand, captures the recommended choice well for analyzing the real brain data in [8], and on the other hand, fulfills the conditions for the asymptotic results. (I) The time-varying stimuli are generated from independent Bernoulli trials such that $P\{s(t_i) = 1\} = 0.5$. (II) Following [9], the HRF is $h(t_j) = g_1(1.5(j-1))/a_1 - g_2(1.5(j-1))/a_2$, $j = 1, \ldots, p_n$, where $g_1(t) = (t - 5.5)^5 \exp\{-(t - 5.5)/0.9\}$ and $g_2(t) = 0.4(t - 5.5)^{12} \exp\{-(t - 5.5)/0.7\}$, $a_1 = \max\{g_1(t)\}$ and $a_2 = \max\{g_2(t)\}$. (III) The drift function is $d(t_i) = \alpha_0 + \alpha_1 t_i + \alpha_2 t_i^2$, $i = 1, \ldots, n$, where $(\alpha_0, \alpha_1, \alpha_2) = (-7.8737, 47.5836, -32.6734)$. (IV) The noise process ϵ is the sum of independent noise processes ϵ_1 and ϵ_2 (see [28]); $\{\epsilon_1(t_i)\}$ are i.i.d. normal with mean zero and variance 0.5216^2, 0.3689^2, 0.2608^2 and 0.1844^2 respectively; ϵ_2 is AR(1), i.e., $\epsilon_2(t_i) = \rho\epsilon_2(t_{i-1}) + z(t_i)$ with $\rho = 0.638$ and $z(t_i)$

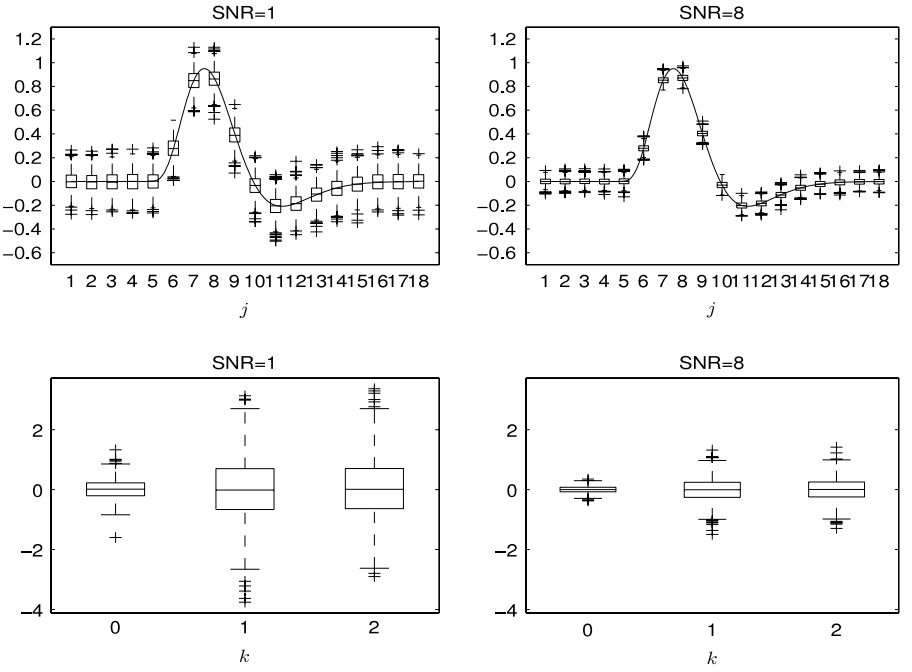

Fig. 2. (Simulated fMRI series in one voxel.) Top panel: boxplots of $\widehat{h}(t_j), j = 1, \ldots, p_n$, where the solid line denotes the true $h(\cdot)$. Bottom panel: boxplots of $\widehat{\alpha}_k - \alpha_k, k = 0, 1, 2$.

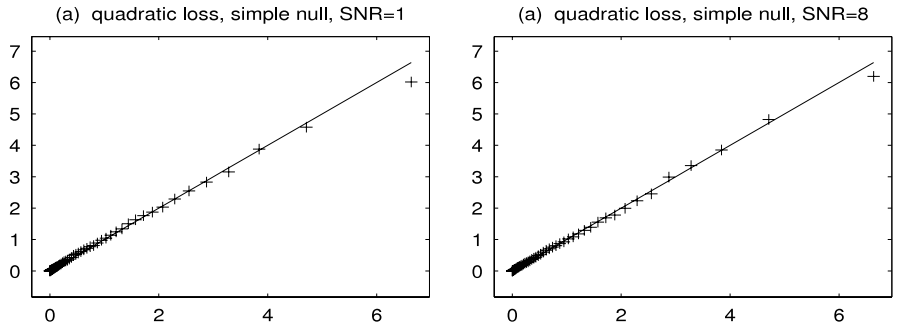

Fig. 3. (Simulated fMRI series in one voxel.) Empirical quantiles (on the y-axis) of test statistics W_n versus quantiles (on the x-axis) of the χ^2_1 distribution. Solid line: the 45° reference line. The null hypothesis is $H_0 : h(t_9) = 0$.

follows the normal distribution with mean zero and variance 0.5216^2, 0.3689^2, 0.2608^2 and 0.1844^2 respectively. These choices give the noise lag-one auto-correlation equal to 0.4 and the signal-to-noise-ratio (SNR) about 1, 2, 4 and 8, where SNR = variance($\mathbf{Sh}$)/variance($\boldsymbol{\epsilon}$). Denote by $\mathrm{cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \sigma^2 R_n$ the error covariance matrix. Following (1.2), the transformed model

$$R_n^{-1/2}\mathbf{y} = R_n^{-1/2}\widetilde{\mathbf{T}}\widetilde{\boldsymbol{\alpha}} + R_n^{-1/2}\mathbf{Sh} + R_n^{-1/2}\boldsymbol{\epsilon} \tag{6.1}$$

is used for estimating $\mathbf{h}$, where the transformed errors are homoscedastic and un-correlated. First, we examine the minimum-BD estimates of the HRF. Fig. 2 displays the boxplots of the HRF estimates $\widehat{h}(t_j)$ along with boxplots of $\widehat{\alpha}_k - \alpha_k$, in which the true R_n is used. Second, we perform the hypothesis testing for $H_0 : h(t_9) = 0$. The fMRI data are simulated in the same way as above except that $\mathbf{h} = \mathbf{0}$ in (1.2). Fig. 3 depicts the QQ plots of the (1st to 99th) percentiles of W_n versus those of χ^2_1. Additionally, Fig. 4 gives the QQ plots for testing $H_0 : h(t_j) = 0, j = 1, \ldots, 18$. Here 18 corresponds to k = 18 in (4.2). It is observed that the Monte Carlo null distribution of W_n could be approximated well by the χ^2 distribution, and that the test under composite null models can be made as precise as the test for the simple null.

In practice, the true error covariance matrix is unknown and needs to be estimated. For computational expedience, we adopt the first-order difference-based method [29] for estimating R_n. The results using the estimated R_n are given in Figs. 5–7, which compare well with counterparts using the true R_n in Figs. 2–4.

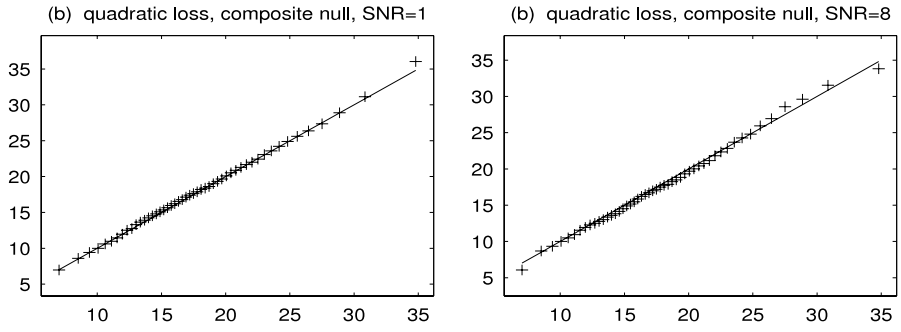

Fig. 4. (Simulated fMRI series in one voxel.) Empirical quantiles (on the y-axis) of test statistics W_n versus quantiles (on the x-axis) of the χ^2_{18} distribution. Solid line: the 45° reference line. The null hypothesis is $H_0 : h(t_j) = 0, j = 1, \ldots, 18$.

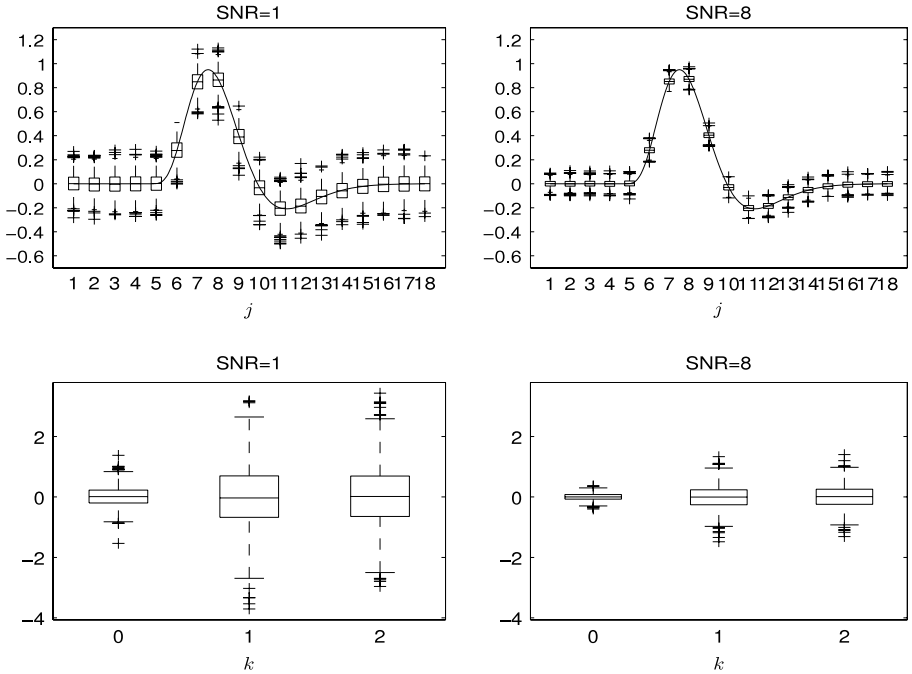

Fig. 5. The captions are similar to those in Fig. 2, except that the estimated R_n is used in (6.1).

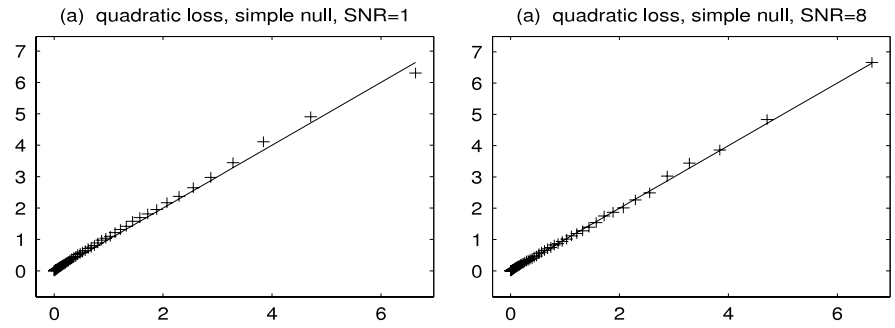

Fig. 6. The captions are similar to those in Fig. 3, except that the estimated R_n is used in (6.1).

6.2. *Impact of* BD *on parametric regression and classification*

To evaluate the impact of loss functions on parametric regression and classification, we conduct a simulation study. We generate data with two-classes from the model,

$$\mathbb{X}_n = (X_1, \ldots, X_{p_n})^T, \quad \{X_j\}_{j=1}^{p_n} \overset{\text{i.i.d.}}{\sim} U(0, 1), \qquad Y_n \mid \mathbb{X}_n = \mathbb{x}_n \sim \text{Bernoulli}\,\{m(\mathbb{x}_n)\},$$

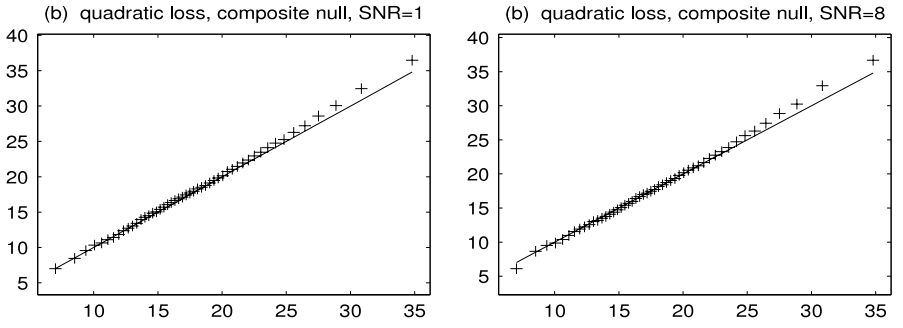

Fig. 7. The captions are similar to those in Fig. 4, except that the estimated R_n is used in (6.1).

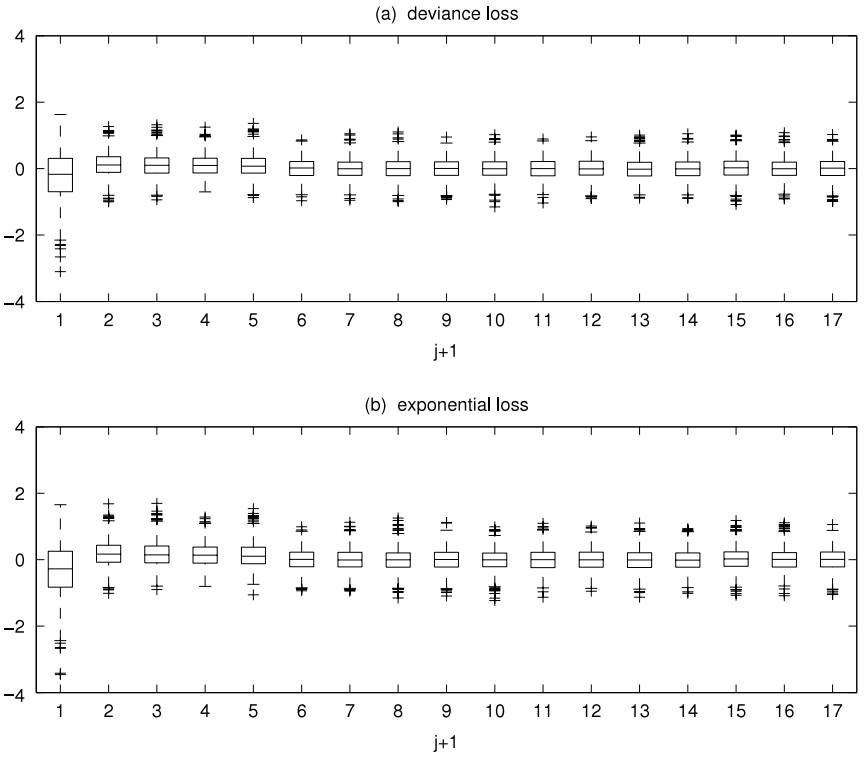

Fig. 8. (Simulated Bernoulli responses.) Boxplots of $\widehat{\beta}_{n,j} - \beta_{n,j;0}, j = 0, 1, \ldots, p_n$ (from left to right in each panel). Panel (a): using the deviance loss; panel (b): using the exponential loss.

where $p_n = [6(n^{1/5.5} - 1)] + 1$,

$$F(m(\mathrm{x}_n)) = \log\left\{\frac{m(\mathrm{x}_n)}{1 - m(\mathrm{x}_n)}\right\} = \widetilde{\mathrm{x}}_n^T \widetilde{\boldsymbol{\beta}}_{n;0}, \tag{6.2}$$

with true values of the parameters $\widetilde{\boldsymbol{\beta}}_{n;0} = (-7.5, 4.5, 4.0, 3.5, 3.0, 0, \ldots, 0)^T$. First, we generate 1000 sets of random samples $\{(\mathrm{X}_{ni}, Y_{ni})\}_{i=1}^n$ of size $n = 1000$ from the distribution of (X_n, Y_n). The minimum-BD estimates $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ are numerically obtained. Fig. 8 compares the boxplots of $\widehat{\beta}_{n,j} - \beta_{n,j;0}, j = 0, 1, \ldots, p_n$, using the deviance loss and the exponential loss. It is observed that the regression estimates under the deviance loss are slightly more centered around the true values with smaller variabilities than those under the exponential loss. This lends support to Proposition 1, since the deviance loss satisfies (3.3) whereas the exponential loss does not. From the regression point of view, the deviance loss does exhibit superiority over the exponential loss in the finite-sample cases. Second, we check the agreement between the asymptotic χ^2 distribution and the finite sampling distribution of W_n under null hypotheses. For simplicity, consider

$$H_0 : \beta_{n,6;0} = 0. \tag{6.3}$$

For each set of the 1000 samplings above, W_n is calculated. The QQ plots of the (1st to 99th) percentiles of W_n against those of the χ_1^2 distribution are displayed in Fig. 9. We observe that the finite sampling null distribution of W_n agrees reasonably

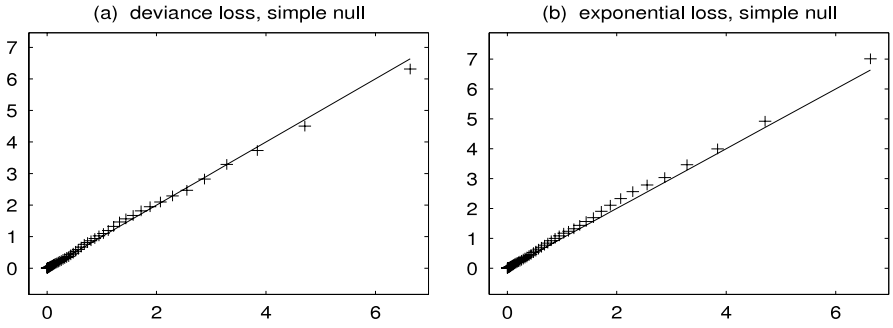

Fig. 9. (Simulated Bernoulli responses.) Empirical quantiles (on the y-axis) of test statistics W_n versus quantiles (on the x-axis) of the χ^2_1 distribution. Solid line: the 45° reference line. The null hypothesis (6.3) is considered.

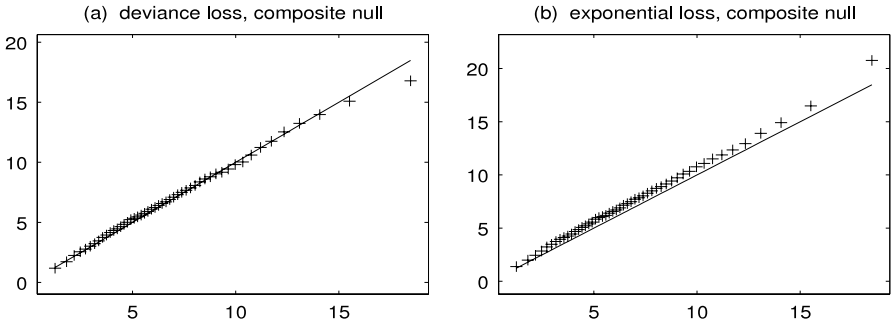

Fig. 10. (Simulated Bernoulli responses.) Empirical quantiles (on the y-axis) of test statistics W_n versus quantiles (on the x-axis) of the χ^2_7 distribution. Solid line: the 45° reference line. The null hypothesis (6.4) is considered.

Table 1
Test misclassification rates using the deviance and exponential losses.

Loss	Test misclassification rates									
Deviance	0.244	0.203	0.204	0.208	0.193	0.185	0.214	0.202	0.226	0.206
Exponential	0.243	0.205	0.206	0.207	0.195	0.186	0.210	0.200	0.219	0.202

well with the χ^2 distribution. This lends support to Theorem 3. Fig. 10 gives analogous results for testing

$$H_0 : \beta_{n,j;0} = 0, \quad j = 9, \ldots, 15. \tag{6.4}$$

Third, we examine the behaviors of classification procedures constructed under different loss functions in the two-class classification. One single training set of size 1000 is used for estimating parameters $\beta_{n,j;0}, j = 0, 1, \ldots, p_n$. Test samples are randomly generated from model (6.2) of size 1000. A comparison of the test misclassification rates in 10 sets of test samples is listed in Table 1. The results indicate that the difference from the deviance and exponential loss functions in regression estimates has a negligible impact on the classification performance. This reinforces the consistency result of Theorem 7.

6.3. Minimum BD estimator for overdispersed count data

In this section, we assess the performance of the minimum-BD estimator when the likelihood of observations is not fully specified. We consider the quasi-likelihood function Q, which relaxes the distributional assumption on a random variable Y via the specification,

$$\partial Q(Y, \mu)/\partial\mu = (Y - \mu)/V(\mu),$$

in which it is assumed that $\mathrm{var}(Y \mid \mathtt{X} = \mathtt{x}) = \sigma^2 V\{E(Y \mid \mathtt{X} = \mathtt{x})\}$ for a nuisance parameter $\sigma^2 > 0$ and a known continuous function $V(\cdot) > 0$. It can be verified that the quasi-likelihood function belongs to BD with the generating q-function as follows,

$$q(\mu) = \int_{-\infty}^{\mu} \frac{s - \mu}{V(s)} \mathrm{d}s.$$

We generate overdispersed Poisson counts Y_{ni} satisfying $\mathrm{var}(Y_{ni} \mid \mathtt{X}_{ni} = \mathtt{x}_{ni}) = 2m(\mathtt{x}_{ni})$, i.e., $V(x) = x$ and the dispersion parameter equal to 2, via a Gamma–Poisson mixture. In the predictor $\mathtt{X}_{ni} = (X_{i1}, X_{i2}, \ldots, X_{ip_n})^T$, $n = 1000$,

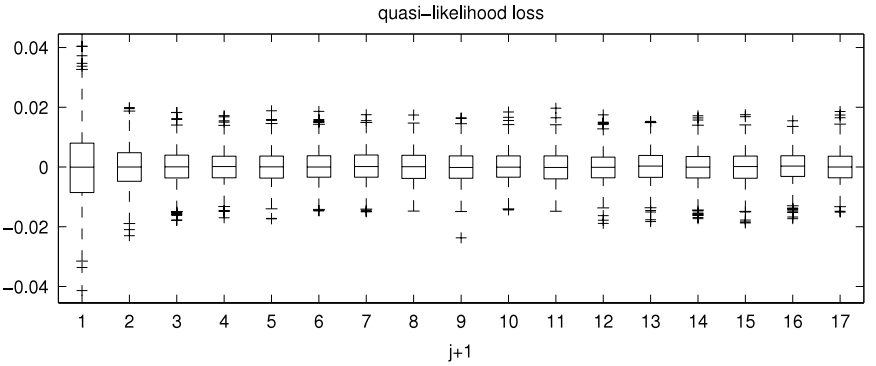

Fig. 11. (Simulated overdispersed Poisson responses.) Boxplots of $\widehat{\beta}_{n,j} - \beta_{n,j;0}, j = 0, 1, \ldots, p_n$ (from left to right).

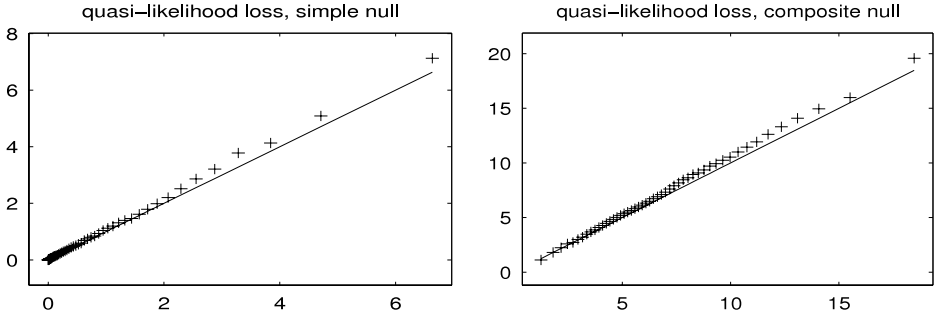

Fig. 12. (Simulated overdispersed Poisson responses.) Empirical quantiles (on the y-axis) of test statistics W_n versus quantiles (on the x-axis) of the χ^2 distribution. Solid line: the 45° reference line. Left panel: for the simple null hypothesis (6.3); right panel: for the composite null hypothesis (6.4).

$p_n = [10(n^{1/5.5} - 1)] + 1, X_{i1} = i/n, (X_{i2}, \ldots, X_{ip_n})$ are i.i.d. Unif $[0, 1]$. The link function is $F(m(\mathrm{x}_n)) = \log\{m(\mathrm{x}_n)\} = \widetilde{\mathrm{x}}_n^T \widetilde{\boldsymbol{\beta}}_{n;0}$, where $\widetilde{\boldsymbol{\beta}}_{n;0} = (3, 4, 2, 0, \ldots, 0)^T$.

Fig. 11 gives the boxplots of $\widehat{\beta}_{n,j} - \beta_{n,j;0}, j = 0, 1, \ldots, p_n$, whereas Fig. 12 gives the QQ-plots of the test statistics W_n for the simple and composite null hypotheses. The conclusions are similar to those in Figs. 8–10.

7. Applications to classification

For binary responses, to evaluate the predictive performance of the minimum-BD classifiers, we randomly split the data into halves, one part for the training set and the other part for the testing set, and calculate the misclassification rate. We replicate this random splitting 200 times and calculate the average of misclassification rates from these 200 runs. Throughout the numerical work in this section, F is set to be the logit link unless otherwise stated.

7.1. Boston housing data

The dataset contains the response MEDV, the median value of owner-occupied homes (in $ 1000's) in 506 US census tracts of the Boston metropolitan area in 1970, along with several explanatory variables which might affect the housing values (see [30]). The covariates CRIM (per capita crime rate by town), ZN (proportion of residential land zoned for lots over 25,000 sq.ft.), INDUS (proportion of non-retail business acres per town), CHAS (Charles River dummy variable (=1 if the tract bounds the river; 0 otherwise)), NOX (nitric oxide concentration (parts per 10 million)), RM (average number of rooms per dwelling), AGE (proportion of owner-occupied units built prior to 1940), DIS (weighted distances to five Boston employment centres), RAD (index of accessibility to radial highways), TAX (full-value property-tax rate per $10,000), PTRATIO (pupil–teacher ratio by town), B (1000 $(\mathrm{Bk} - 0.63)^2$ where Bk is the proportion of blacks by town) and LSTAT (% lower status of the population) are denoted by $X_1, \ldots, X_{13}$ respectively.

To predict whether the median value of owner-occupied homes, denoted by Y, can be categorized as either "high" or "low" (compared with the average of MEDV), the logistic regression model

$$\mathrm{logit}\{P(Y^* = 1 \mid X_1 = x_1, \ldots, X_{13} = x_{13})\} = \beta_0 + \sum_{j=1}^{13} \beta_j x_j$$

Table 2
Classification of Boston housing data.

Loss	Average misclassification rate	
	Using all variables	Using significant variables
Deviance	0.1346	0.1270
Exponential	0.1422	0.1339

is fitted to the data set, where Y^* equals 1 if Y exceeds the average of MEDV and 0 otherwise. Table 2 indicates that the Average Misclassification Rate is similar under the deviance loss and the exponential loss. Interestingly, applying the generalized Wald type test (at level 0.05) proposed in Section 4, variables DIS, RAD, TAX, PTRATIO, LSTAT each are significant using the deviance loss, whereas variables RM, AGE, DIS, RAD, TAX, PTRATIO, LSTAT each are significant using the exponential loss. Identifying a smaller subset of significant input variables appears to be more effective in reducing the misclassification error.

8. Discussion

Stochastic modeling of large-scale datasets usually involves a varying-dimensional model space. This paper concerns statistical regression estimation and inference when the dimension p_n diverges with n, and the loss function Q belongs to a wide class of BD, particularly useful and flexible in situations where the full likelihood is unknown or incompletely specified. Our study reveals that under mild regularity conditions on p_n and Q, the asymptotic distribution of the minimum-BD estimator relies on the loss function only through the second derivative of its generating q-function. If q satisfies the "*generalized Bartlett identity*", then the asymptotic covariance matrix of the estimator achieves the lower bound. The inference procedure for the minimum-BD estimator is also carefully developed. Moreover, we show that though the choice of loss function affects the regression estimation procedure, it has an asymptotically relatively negligible impact on the classification's performance.

A few issues need to be discussed. First, if the loss is quadratic and the link is the identity link, then the rate of p_n can be relaxed from $p_n^4/n = o(1)$ or $p_n^5/n = o(1)$ to $p_n^3/n = o(1)$ without violating the asymptotic results in the paper. Second, for many high-dimensional models, the number of relevant variables is far fewer than the sample size, i.e. the signal is "sparse". In this case, it is more efficient to carry out variable selection and dimension reduction techniques, followed by statistical inference developed in this paper for models with the subset of p_n contributive variables. Third, the penalized minimum-BD estimators and classifiers for models with dimension p_n of the order n or larger are potentially effective for selecting important variables, and we plan to report the details in future work.

Acknowledgments

The author thanks the Associate Editor and two referees for insightful comments and suggestions. The research was supported by National Science Foundation grants.

Appendix. Proofs of main results

For a matrix M, its eigenvalues, minimum eigenvalue, maximum eigenvalue and trace are labeled by $\lambda_j(M)$, $\lambda_{\min}(M)$, $\lambda_{\max}(M)$ and $\mathrm{tr}(M)$ respectively. Let $\|M\| = \sup\{\|M\mathrm{x}\| : \|\mathrm{x}\| = 1\} = \{\lambda_{\max}(M^TM)\}^{1/2}$ be the matrix L_2 norm, which for symmetric matrices reduces to $\|M\| = \max_j\{|\lambda_j(M)|\}$. The Frobenius norm of a matrix M is $\|M\|_F = \{\mathrm{tr}(M^TM)\}^{1/2}$. See [31] for details. Throughout the proof, C is used as a generic finite constant.

We first impose some technical conditions, which are not the weakest possible but facilitate the technical derivations.
Condition A:

A1. $\sup_{n\geq 1}\|\widetilde{\boldsymbol{\beta}}_{n;0}\|_1 < \infty$ and $\sup_{n\geq 1}\|\mathrm{X}_n\|_\infty < \infty$.
A2. $E(\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T)$ exists and is nonsingular.
A3. $\sup_{n\geq 1}E(Y_n^2) < \infty$.
A4. There is a large enough open subset of $\mathbb{R}^{p_n+1}$ which contains the true parameter point $\widetilde{\boldsymbol{\beta}}_{n;0}$, such that $F^{-1}(\widetilde{\mathrm{X}}_n^T\widetilde{\boldsymbol{\beta}}_n)$ is bounded for all $\widetilde{\boldsymbol{\beta}}_n$ in the subset.
A5. The eigenvalues of $\mathbf{H}_n = -E[q^{(2)}(m(\mathrm{X}_n))/\{F^{(1)}(m(\mathrm{X}_n))\}^2\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T]$ are uniformly bounded away from 0.
A6. $q^{(4)}(\cdot)$ is continuous, and $q^{(2)}(\cdot) < 0$.
A7. $F(\cdot)$ is a bijection, $F^{(3)}(\cdot)$ is continuous, and $F^{(1)}(\cdot) \neq 0$.

Condition B: is identical to Condition A except that A3 and A5 are replaced by B3 and B5 below respectively.

B3. There exists some $\delta \geq 1/2$ such that $\sup_{n\geq 1}E(|Y_n|^{2+\delta}) < \infty$.
B5. The eigenvalues of Ω_n and $\mathbf{H}_n$ are uniformly bounded away from 0. Also, $\|\mathbf{H}_n^{-1}\Omega_n\|$ is bounded away from ∞.

Condition C: is identical to Condition B except that B4 is replaced by C4 below.

C4. There is a large enough open subset of $\mathbb{R}^{p_n+1}$ which contains the true parameter point $\widetilde{\boldsymbol{\beta}}_{n;0}$, such that $A_n\widetilde{\boldsymbol{\beta}}_{n;0} = \mathbf{0}$, and $F^{-1}(\widetilde{\mathrm{X}}_n^T\widetilde{\boldsymbol{\beta}}_n)$ is bounded for all $\widetilde{\boldsymbol{\beta}}_n$ in the subset.

Condition D: is identical to Condition C except that C5 is replaced by D5 below.

D5. The eigenvalues of $\mathbf{H}_n$ are uniformly bounded away from 0. Also, $\|\mathbf{H}_n^{-1/2}\Omega_n^{1/2}\|$ is bounded away from ∞.

Proof of Theorem 1

Let $r_n = o(1)$ (whose exact order will be chosen later) and $\widetilde{\mathbf{u}}_n = (u_0, u_1, \ldots, u_{p_n})^T \in \mathbb{R}^{p_n+1}$. Following the idea of the proof in [18], it suffices to show that for any given $\epsilon > 0$, there is a sufficiently large constant C_ϵ such that, for large n we have

$$P\left\{\inf_{\|\widetilde{\mathbf{u}}_n\|=C_\epsilon} \ell_n(\widetilde{\boldsymbol{\beta}}_{n;0} + r_n\widetilde{\mathbf{u}}_n) > \ell_n(\widetilde{\boldsymbol{\beta}}_{n;0})\right\} \geq 1-\epsilon. \tag{A.1}$$

This implies that with probability at least $1-\epsilon$, there exists a local minimizer $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ of $\ell_n(\widetilde{\boldsymbol{\beta}}_n)$ in the ball $\{\widetilde{\boldsymbol{\beta}}_{n;0}+r_n\widetilde{\mathbf{u}}_n : \|\widetilde{\mathbf{u}}_n\| \leq C_\epsilon\}$ such that $\|\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}\| = O_P(r_n)$. To show (A.1), consider

$$\ell_n(\widetilde{\boldsymbol{\beta}}_{n;0} + r_n\widetilde{\mathbf{u}}_n) - \ell_n(\widetilde{\boldsymbol{\beta}}_{n;0}) = \frac{1}{n}\sum_{i=1}^n\{Q(Y_{ni}, F^{-1}(\widetilde{\mathrm{X}}_{ni}^T(\widetilde{\boldsymbol{\beta}}_{n;0} + r_n\widetilde{\mathbf{u}}_n))) - Q(Y_{ni}, F^{-1}(\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0}))\} \equiv I_1, \tag{A.2}$$

where $\|\widetilde{\mathbf{u}}_n\| = C_\epsilon$.

Then from (3.1), for $\mu = F^{-1}(\theta)$,

$$\begin{aligned}
&\mathrm{q}_1(y;\theta) = (y-\mu)q^{(2)}(\mu)/F^{(1)}(\mu),\\
&\mathrm{q}_2(y;\theta) = -q^{(2)}(\mu)/\{F^{(1)}(\mu)\}^2 + (y-\mu)A_1(\mu),\\
&\mathrm{q}_3(y;\theta) \equiv A_2(\mu) + (y-\mu)A_3(\mu),
\end{aligned} \tag{A.3}$$

where $A_1(\mu) = \{q^{(3)}F^{(1)} - q^{(2)}F^{(2)}\}/\{F^{(1)}\}^3(\mu)$, $A_2(\mu) = \{-2q^{(3)}F^{(1)} + 3q^{(2)}F^{(2)}\}/\{F^{(1)}\}^4(\mu)$ and $A_3(\mu) = [q^{(4)}\{F^{(1)}\}^2 - 3q^{(3)}F^{(1)}F^{(2)} - q^{(2)}F^{(1)}F^{(3)} + 3q^{(2)}\{F^{(2)}\}^2]/\{F^{(1)}\}^5(\mu)$. By Taylor's expansion,

$$I_1 = I_{1,1} + I_{1,2} + I_{1,3}, \tag{A.4}$$

where

$$\begin{aligned}
I_{1,1} &= \frac{r_n}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}^T\widetilde{\mathbf{u}}_n,\\
I_{1,2} &= \frac{r_n^2}{2n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})(\widetilde{\mathrm{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^2,\\
I_{1,3} &= \frac{r_n^3}{6n}\sum_{i=1}^n \mathrm{q}_3(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)(\widetilde{\mathrm{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^3
\end{aligned}$$

for $\widetilde{\boldsymbol{\beta}}_n^*$ located between $\widetilde{\boldsymbol{\beta}}_{n;0}$ and $\widetilde{\boldsymbol{\beta}}_{n;0} + r_n\widetilde{\mathbf{u}}_n$. Hence

$$|I_{1,1}| \leq r_n\left\|\frac{1}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}\right\| \|\widetilde{\mathbf{u}}_n\| = O_P(r_n\sqrt{p_n/n})\|\widetilde{\mathbf{u}}_n\|. \tag{A.5}$$

For $I_{1,2}$ in (A.4), Eq. (A.3) gives that

$$\begin{aligned}
I_{1,2} &= -\frac{r_n^2}{2n}\sum_{i=1}^n \frac{q^{(2)}(m(\mathrm{X}_{ni}))}{\{F^{(1)}(m(\mathrm{X}_{ni}))\}^2}(\widetilde{\mathrm{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^2 + \frac{r_n^2}{2n}\sum_{i=1}^n\{Y_{ni} - m(\mathrm{X}_{ni})\}A_1(m(\mathrm{X}_{ni}))(\widetilde{\mathrm{X}}_{ni}^T\widetilde{\mathbf{u}}_n)^2\\
&\equiv I_{1,2,1} + I_{1,2,2}.
\end{aligned}$$

Note that

$$\left\|\frac{1}{n}\sum_{i=1}^n \frac{q^{(2)}(m(\mathrm{X}_{ni}))}{\{F^{(1)}(m(\mathrm{X}_{ni}))\}^2}\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T - E\left[\frac{q^{(2)}(m(\mathrm{X}_n))}{\{F^{(1)}(m(\mathrm{X}_n))\}^2}\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T\right]\right\|_F = O_P(p_n/\sqrt{n}).$$

Thus

$$I_{1,2,1} = -\frac{r_n^2}{2}\widetilde{\mathbf{u}}_n^T E\left[\frac{q^{(2)}(m(\mathrm{X}_n))}{\{F^{(1)}(m(\mathrm{X}_n))\}^2}\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T\right]\widetilde{\mathbf{u}}_n + r_n^2 O_P(p_n/\sqrt{n})\|\widetilde{\mathbf{u}}_n\|^2.$$

Meanwhile, we have

$$|I_{1,2,2}| \le r_n^2\left\|\frac{1}{n}\sum_{i=1}^n\{Y_{ni} - m(\mathrm{X}_{ni})\}A_1(m(\mathrm{X}_{ni}))\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T\right\|_F\|\widetilde{\mathbf{u}}_n\|^2 = r_n^2 O_P(p_n/\sqrt{n})\|\widetilde{\mathbf{u}}_n\|^2.$$

Thus,

$$I_{1,2} = -\frac{r_n^2}{2}\widetilde{\mathbf{u}}_n^T E\left[\frac{q^{(2)}(m(\mathrm{X}_n))}{\{F^{(1)}(m(\mathrm{X}_n))\}^2}\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T\right]\widetilde{\mathbf{u}}_n + O_P(r_n^2 p_n/\sqrt{n})\|\widetilde{\mathbf{u}}_n\|^2.$$

For $I_{1,3}$ in (A.4), we observe that

$$|I_{1,3}| \le r_n^3\frac{1}{n}\sum_{i=1}^n|\mathrm{q}_3(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)||\widetilde{\mathrm{X}}_{ni}^T\widetilde{\mathbf{u}}_n|^3 = O_P(r_n^3 p_n^{3/2})\|\widetilde{\mathbf{u}}_n\|^3,$$

which follows from Conditions A1 and A4.

By setting $r_n = \sqrt{p_n/n}$, using (A.5) and $p_n^4/n \to 0$, we can choose C_ϵ large enough such that both $I_{1,1}$ and $I_{1,3}$ are dominated by the first term of $I_{1,2}$, which is positive by Condition A5. This in turn implies (A.1). □

Proof of Theorem 2

Notice the estimating equations $\frac{\partial\ell_n(\widetilde{\boldsymbol{\beta}}_n)}{\partial\boldsymbol{\beta}_n}|_{\widetilde{\boldsymbol{\beta}}_n=\widehat{\widetilde{\boldsymbol{\beta}}}_n} = \mathbf{0}$, since $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ is a local minimizer of $\ell_n(\widetilde{\boldsymbol{\beta}}_n)$. Taylor's expansion applied to the left side of the estimating equations yields

$$\begin{aligned}\mathbf{0} &= \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}\right\} + \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T\right\}(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0})\\ &\quad + \frac{1}{2n}\sum_{i=1}^n \mathrm{q}_3(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)\{\widetilde{\mathrm{X}}_{ni}^T(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0})\}^2\widetilde{\mathrm{X}}_{ni}\\ &\equiv \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}\right\} + K_2(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) + K_3,\end{aligned} \tag{A.6}$$

where $\widetilde{\boldsymbol{\beta}}_n^*$ lies between $\widetilde{\boldsymbol{\beta}}_{n;0}$ and $\widehat{\widetilde{\boldsymbol{\beta}}}_n$. Below, we will show

$$\|K_2 - \mathbf{H}_n\| = O_P(p_n/\sqrt{n}), \tag{A.7}$$

$$\|K_3\| = O_P(p_n^{5/2}/n). \tag{A.8}$$

First, to show (A.7), note that $K_2 - \mathbf{H}_n \equiv L_1$, where

$$\begin{aligned}L_1 &= -\left(\frac{1}{n}\sum_{i=1}^n\frac{q^{(2)}(m(\mathrm{X}_{ni}))}{\{F^{(1)}(m(\mathrm{X}_{ni}))\}^2}\widetilde{\mathrm{X}}_{ni}^T\widetilde{\mathrm{X}}_{ni}^T - E\left[\frac{q^{(2)}(m(\mathrm{X}_n))}{\{F^{(1)}(m(\mathrm{X}_n))\}^2}\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T\right]\right) + \frac{1}{n}\sum_{i=1}^n\{Y_{ni} - m(\mathrm{X}_{ni})\}A_1(m(\mathrm{X}_{ni}))\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T\\ &\equiv L_{1,1} + L_{1,2}.\end{aligned}$$

Similar arguments for the proof of Theorem 1 give $\|L_{1,1}\| = O_P(p_n/\sqrt{n})$ and $\|L_{1,2}\| = O_P(p_n/\sqrt{n})$. Thus $\|L_1\| = O_P(p_n/\sqrt{n})$.

Second, $\|K_3\| \le O_P(p_n^{3/2})O_P(p_n/n)$ completes (A.8).

Third, by (A.6)–(A.8) and $\|\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}\| = O_P(\sqrt{p_n/n})$, we see that

$$\mathbf{H}_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) = -\frac{1}{n}\sum_{i=1}^n \mathrm{q}_1(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni} + \mathbf{u}_n, \tag{A.9}$$

where $\|\mathbf{u}_n\| = O_P(p_n^{5/2}/n)$. Note that by Condition B5,

$$\|\sqrt{n}A_n\Omega_n^{-1/2}\mathbf{u}_n\| \le \sqrt{n}\|A_n\|_F\lambda_{\max}(\Omega_n^{-1/2})\|\mathbf{u}_n\| = O_P(p_n^{5/2}/\sqrt{n}) = o_P(1).$$

Thus

$$\sqrt{n}A_n\Omega_n^{-1/2}\mathbf{H}_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) = -\frac{1}{\sqrt{n}}A_n\Omega_n^{-1/2}\sum_{i=1}^n \mathrm{q}_1(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni} + o_P(1).$$

To complete proving Theorem 2, we apply the Lindeberg–Feller central limit theorem [21] to $\sum_{i=1}^n \boldsymbol{Z}_{ni}$, where $\boldsymbol{Z}_{ni} = -n^{-1/2}A_n\Omega_n^{-1/2}\mathrm{q}_1(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}$. It suffices to check (I) $\sum_{i=1}^n \mathrm{cov}(\boldsymbol{Z}_{ni}) \to G$ and (II) $\sum_{i=1}^n E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) = o(1)$ for some $\delta > 0$. Condition (I) follows from $\mathrm{var}\{\mathrm{q}_1(Y_n;\widetilde{\mathrm{X}}_n^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_n\} = \Omega_n$. To verify condition (II), notice that

$$\begin{aligned}
E(\|\boldsymbol{Z}_n\|^{2+\delta}) &\le n^{-(2+\delta)/2}E\left\{\|A_n\|_F^{2+\delta}\left[\|\Omega_n^{-1/2}\widetilde{\mathrm{X}}_n\|\left|\frac{q^{(2)}(m(\mathrm{X}_n))}{F^{(1)}(m(\mathrm{X}_n))}\{Y_n - m(\mathrm{X}_n)\}\right|\right]^{2+\delta}\right\}\\
&\le Cn^{-(2+\delta)/2}E[\{\lambda_{\min}^{-1/2}(\Omega_n)\|\widetilde{\mathrm{X}}_n\|\}^{2+\delta}|Y_n - m(\mathrm{X}_n)|^{2+\delta}]\\
&\le Cp_n^{(2+\delta)/2}n^{-(2+\delta)/2}E\{|Y_n - m(\mathrm{X}_n)|^{2+\delta}\}\\
&\le 2Cp_n^{(2+\delta)/2}n^{-(2+\delta)/2}[E(|Y_n|^{2+\delta}) + E\{|m(\mathrm{X}_n)|^{2+\delta}\}]\\
&= O\{(p_n/n)^{(2+\delta)/2}\}.
\end{aligned}$$

Thus, we get $\sum_{i=1}^n E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) \le O\{n(p_n/n)^{(2+\delta)/2}\} = O\{p_n^{(2+\delta)/2}/n^{\delta/2}\}$, which is $o(1)$ by Condition B3. This verifies Condition (II). □

Proof of Proposition 1

The proof follows from the result: "For appropriately dimensioned random matrices A and B, if $E(BB^T)$ is positive definite, then $E(AA^T) \ge E(AB^T)\{E(BB^T)\}^{-1}E(BA^T)$. Moreover, if $B = cA$ for a constant $c \ne 0$, then the inequality becomes an equality". □

Proof of Proposition 2

Note $\|A_n(\widehat{V}_n - V_n)A_n^T\| \le \|\widehat{V}_n - V_n\|\,\|A_n\|_F^2$. Since $\|A_n\|_F^2 \to \mathrm{tr}(G)$, it suffices to prove that $\|\widehat{V}_n - V_n\| = o_P(1)$.

First, we prove $\|\widehat{\mathbf{H}}_n - \mathbf{H}_n\| = o_P(1)$. Note that

$$\begin{aligned}
\widehat{\mathbf{H}}_n - \mathbf{H}_n &= \frac{1}{n}\sum_{i=1}^n\{\mathrm{q}_2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}}_n) - \mathrm{q}_2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\}\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T + \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T - \mathbf{H}_n\right\}\\
&\equiv I_1 + I_2.
\end{aligned}$$

From the proof of (A.7) in Theorem 2, we know that $\|I_2\| = O_P(p_n/\sqrt{n}) = o_P(1)$. We only need to consider the term I_1,

$$\begin{aligned}
I_1 &= -\frac{1}{n}\sum_{i=1}^n\left[\frac{q^{(2)}(\widehat{m}(\mathrm{X}_{ni}))}{\{F^{(1)}(\widehat{m}(\mathrm{X}_{ni}))\}^2} - \frac{q^{(2)}(m(\mathrm{X}_{ni}))}{\{F^{(1)}(m(\mathrm{X}_{ni}))\}^2}\right]\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T\\
&\quad + \frac{1}{n}\sum_{i=1}^n[\{Y_{ni} - \widehat{m}(\mathrm{X}_{ni})\}A_1(\widehat{m}(\mathrm{X}_{ni})) - \{Y_{ni} - m(\mathrm{X}_{ni})\}A_1(m(\mathrm{X}_{ni}))]\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T\\
&\equiv I_{1,1} + I_{1,2}.
\end{aligned}$$

Let $g(\cdot) = q^{(2)}(\cdot)/\{F^{(1)}(\cdot)\}^2$. By the assumptions, $g(\cdot)$ is differentiable. Thus

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n|g(\widehat{m}(\mathrm{X}_{ni})) - g(m(\mathrm{X}_{ni}))| &= \frac{1}{n}\sum_{i=1}^n|(g\circ F^{-1})'(\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)\mathrm{X}_{ni}^T(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0})|\\
&= O_P(1)O_P(\sqrt{p_n})O_P(\sqrt{p_n/n}) = O_P(p_n/\sqrt{n}),
\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}_n^*$ is between $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ and $\widetilde{\boldsymbol{\beta}}_{n;0}$. Thus

$$\|I_{1,1}\| \le \left\|\frac{1}{n}\sum_{i=1}^n|g(\widehat{m}(\mathrm{X}_{ni})) - g(m(\mathrm{X}_{ni}))|\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T\right\|_F = O_P(p_n/\sqrt{n})O_P(p_n) = O_P(p_n^2/\sqrt{n}).$$

Similar arguments give $\|I_{1,2}\| = O_P(p_n^2/\sqrt{n})$. Thus $\|I_1\| = O_P(p_n^2/\sqrt{n}) = o_P(1)$.

Second, we show $\|\widehat{\Omega}_n - \Omega_n\| = o_P(1)$. It is easy to see that

$$\begin{aligned}
\widehat{\Omega}_n - \Omega_n &= \frac{1}{n}\sum_{i=1}^n\{\mathrm{q}_1^2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}}_n) - \mathrm{q}_1^2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\}\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T + \left\{\frac{1}{n}\sum_{i=1}^n \mathrm{q}_1^2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T - \Omega_n\right\}\\
&= \Delta_{1,1} + \Delta_{1,2},
\end{aligned}$$

where $\|\Delta_{1,1}\| = O_P(p_n^2/\sqrt{n})$ and $\|\Delta_{1,2}\| = O_P(p_n/\sqrt{n})$. We observe that $\|\widehat{\Omega}_n - \Omega_n\| = O_P(p_n^2/\sqrt{n}) = o_P(1)$.

Third, we show $\|\widehat{V}_n - V_n\| = o_P(1)$. Note $\widehat{V}_n - V_n = L_1 + L_2 + L_3$, where $L_1 = \widehat{\mathbf{H}}_n^{-1}(\widehat{\Omega}_n - \Omega_n)\widehat{\mathbf{H}}_n^{-1}$, $L_2 = \widehat{\mathbf{H}}_n^{-1}(\mathbf{H}_n - \widehat{\mathbf{H}}_n)\mathbf{H}_n^{-1}\Omega_n\widehat{\mathbf{H}}_n^{-1}$ and $L_3 = \mathbf{H}_n^{-1}\Omega_n\widehat{\mathbf{H}}_n^{-1}(\mathbf{H}_n - \widehat{\mathbf{H}}_n)\mathbf{H}_n^{-1}$. By Assumption B5, it is straightforward to verify that $\|\mathbf{H}_n^{-1}\| \le O(1)$, $\|\widehat{\mathbf{H}}_n^{-1}\| \le O_P(1)$ and $\|\mathbf{H}_n^{-1}\Omega_n\| \le O(1)$. Since $\|L_1\| \le \|\widehat{\mathbf{H}}_n^{-1}\|\,\|\widehat{\Omega}_n - \Omega_n\|\,\|\widehat{\mathbf{H}}_n^{-1}\|$, we conclude $\|L_1\| = o_P(1)$, and similarly $\|L_2\| = o_P(1)$ and $\|L_3\| = o_P(1)$. Hence $\widehat{V}_n - V_n = o_P(1)$. □

Proof of Theorem 3

Before showing Theorem 3, Lemma 1 is needed.

Lemma 1. *Assume the conditions of Theorem 3. Then*

$$\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0} = -\frac{1}{n}\mathbf{H}_n^{-1}\sum_{i=1}^{n} \mathrm{q}_1(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni} + o_P(n^{-1/2}),$$

$$\sqrt{n}(A_n\widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1}A_n^T)^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{\mathrm{k}}).$$

Proof. Following (A.9) in the proof of Theorem 2, we observe that $\|\mathbf{u}_n\| = O_P(p_n^{5/2}/n) = o_P(n^{-1/2})$. Condition B5 completes the proof for the first part.

To show the second part, denote $U_n = A_n\mathbf{H}_n^{-1}\Omega_n\mathbf{H}_n^{-1}A_n^T = A_nV_nA_n^T$ and $\widehat{U}_n = A_n\widehat{\mathbf{H}}_n^{-1}\widehat{\Omega}_n\widehat{\mathbf{H}}_n^{-1}A_n^T = A_n\widehat{V}_nA_n^T$. Notice that the eigenvalues of V_n are uniformly bounded away from 0. So are the eigenvalues of U_n. From the first part, we see that

$$A_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) = -\frac{1}{n}A_n\mathbf{H}_n^{-1}\sum_{i=1}^{n} \mathrm{q}_1(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni} + o_P(n^{-1/2}).$$

It follows that $\sqrt{n}U_n^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) = \sum_{i=1}^n \boldsymbol{Z}_{ni} + o_P(1)$, where $\boldsymbol{Z}_{ni} = -n^{-1/2}U_n^{-1/2}A_n\mathbf{H}_n^{-1}\mathrm{q}_1(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}$. To show $\sum_{i=1}^n \boldsymbol{Z}_{ni} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{\mathrm{k}})$, similar to the proof for Theorem 2, we check (III) $\sum_{i=1}^n \mathrm{cov}(\boldsymbol{Z}_{ni}) \to \mathbf{I}_{\mathrm{k}}$ and (IV) $\sum_{i=1}^n E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) = o(1)$ for some $\delta > 0$. Condition (III) is straightforward since $\sum_{i=1}^n \mathrm{cov}(\boldsymbol{Z}_{ni}) = U_n^{-1/2}U_nU_n^{-1/2} = \mathbf{I}_{\mathrm{k}}$. To check condition (IV), we see that $E(\|\boldsymbol{Z}_n\|^{2+\delta}) = O\{(p_n/n)^{(2+\delta)/2}\}$. This and Condition B3 yield $\sum_{i=1}^n E(\|\boldsymbol{Z}_{ni}\|^{2+\delta}) \le O\{p_n^{(2+\delta)/2}/n^{\delta/2}\} = o(1)$. Hence

$$\sqrt{n}U_n^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{\mathrm{k}}). \tag{A.10}$$

From the proof of Proposition 2, it can be concluded that $\|\widehat{U}_n - U_n\| = o_P(1)$ and that the eigenvalues of $\widehat{U}_n$ are uniformly bounded away from 0 and ∞ with a probability tending to one. Consequently,

$$\|\widehat{U}_n^{-1/2}U_n^{1/2} - \mathbf{I}_{\mathrm{k}}\| = o_P(1). \tag{A.11}$$

Combining (A.10) and (A.11) and Slutsky's Theorem completes the proof that $\sqrt{n}\widehat{U}_n^{-1/2}A_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{\mathrm{k}})$. □

We now show Theorem 3, which follows directly from H_0 in (4.1) and the second part of Lemma 1. □

Proof of Theorem 4

For the matrix A_n in (4.1), there exists a $(p_n + 1 - \mathrm{k}) \times (p_n + 1)$ matrix B_n satisfying $B_nB_n^T = \mathbf{I}_{p_n+1-\mathrm{k}}$ and $A_nB_n^T = \mathbf{0}$. Therefore, $A_n\widetilde{\boldsymbol{\beta}}_n = \mathbf{0}$ is equivalent to $\widetilde{\boldsymbol{\beta}}_n = B_n^T\boldsymbol{\gamma}_n$, where $\boldsymbol{\gamma}_n$ is a $(p_n + 1 - \mathrm{k}) \times 1$ vector. Thus under H_0 in (4.1), we have $\widetilde{\boldsymbol{\beta}}_{n;0} = B_n^T\boldsymbol{\gamma}_{n;0}$. Then minimizing $\ell_n(\widetilde{\boldsymbol{\beta}}_n)$ subject to $A_n\widetilde{\boldsymbol{\beta}}_n = \mathbf{0}$ is equivalent to minimizing $\ell_n(B_n^T\boldsymbol{\gamma}_n)$ with respect to $\boldsymbol{\gamma}_n$, and we denote by $\widehat{\boldsymbol{\gamma}}_n$ the minimizer. Note that under (3.2), $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ is the unique minimizer of $\ell_n(\widetilde{\boldsymbol{\beta}}_n)$. Hence $\Lambda_n = 2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n)\}$. Before showing Theorem 4, we need Lemma 2.

Lemma 2. *Assume conditions of Theorem 4. Then under H_0 in (4.1), we have that* $B_n^T(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n;0}) = -n^{-1}B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}B_n\sum_{i=1}^n \mathrm{q}_1(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni} + o_P(n^{-1/2})$, *and* $2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n)\} = n(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n)^T\mathbf{H}_n(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n) + o_P(1)$.

Proof. To obtain the first part, following the proof of (A.9) in Theorem 2, we have a similar expression for $\widehat{\boldsymbol{\gamma}}_n$,

$$B_n\mathbf{H}_nB_n^T(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n;0}) = -\frac{1}{n}B_n\sum_{i=1}^{n} \mathrm{q}_1(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni} + \mathbf{w}_n,$$

with $\|\mathbf{w}_n\| = o_P(n^{-1/2})$. As a result,

$$B_n^T(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_{n;0}) = -\frac{1}{n}B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}B_n\sum_{i=1}^{n} \mathrm{q}_1(Y_{ni}; \widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni} + B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}\mathbf{w}_n.$$

We notice that

$$\|B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}\mathbf{w}_n\| \le \|(B_n\mathbf{H}_nB_n^T)^{-1}\|\,\|\mathbf{w}_n\| \le \|\mathbf{w}_n\|/\lambda_{\min}(\mathbf{H}_n) = o_P(n^{-1/2}),$$

in which the fact $\lambda_{\min}(B_n\mathbf{H}_nB_n^T) \ge \lambda_{\min}(\mathbf{H}_n)$ is used.

The proof of the second part proceed in three steps. In Step 1, we use the following Taylor expansion for $\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n)$,

$$\begin{aligned}\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n) &= \frac{1}{2n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}}_n)\{\widetilde{\mathrm{X}}_{ni}^T(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n)\}^2 + \frac{1}{6n}\sum_{i=1}^n \mathrm{q}_3(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_n^*)\{\widetilde{\mathrm{X}}_{ni}^T(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n)\}^3\\ &\equiv I_1 + I_2,\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}_n^*$ lies between $\widehat{\widetilde{\boldsymbol{\beta}}}_n$ and $B_n^T\widehat{\boldsymbol{\gamma}}_n$.

In Step 2, we analyze the stochastic order of $B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n$. For a matrix X whose column vectors are linearly independent, set $P_X = X(X^TX)^{-1}X^T$. Define $H_n = \mathbf{I}_{p_n+1} - P_{\mathbf{H}_n^{1/2}B_n^T}$. Then $\mathbf{H}_n^{-1} - B_n^T(B_n\mathbf{H}_nB_n^T)^{-1}B_n = \mathbf{H}_n^{-1/2}H_n\mathbf{H}_n^{-1/2}$. By Lemma 1 and the first part of Lemma 2, we see immediately that

$$B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n = \mathbf{H}_n^{-1/2}H_n\mathbf{H}_n^{-1/2}\left(\frac{1}{n}\sum_{i=1}^n \mathrm{q}_{1,i}\widetilde{\mathrm{X}}_{ni}\right) + o_P(n^{-1/2}), \tag{A.12}$$

where $\mathrm{q}_{1,i} = \mathrm{q}_1(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})$. Note that $\|\mathbf{H}_n^{-1/2}H_n\mathbf{H}_n^{-1/2}(n^{-1}\sum_{i=1}^n \mathrm{q}_{1,i}\widetilde{\mathrm{X}}_{ni})\| = O_P(1/\sqrt{n})$. This gives

$$\|B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n\| = O_P(1/\sqrt{n}). \tag{A.13}$$

In Step 3, we conclude from (A.13) that $I_2 = O_P\{(p_n/n)^{3/2}\} = o_P(1/n)$. Then $2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n, \mathbf{0})\} = 2nI_1 + o_P(1)$. Similar to the proof of Proposition 2, it is straightforward to see that

$$\begin{aligned}2nI_1 &= n(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n)^T\left\{\frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widehat{\widetilde{\boldsymbol{\beta}}}_n)\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T\right\}(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n)\\ &= n(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n)^T\left\{\frac{1}{n}\sum_{i=1}^n \mathrm{q}_2(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni}\widetilde{\mathrm{X}}_{ni}^T\right\}(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n) + o_P(1)\\ &= n(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n)^T E\{\mathrm{q}_2(Y_n;\widetilde{\mathrm{X}}_n^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_n\widetilde{\mathrm{X}}_n^T\}(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n) + o_P(1)\\ &= n(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n)^T\mathbf{H}_n(B_n^T\widehat{\boldsymbol{\gamma}}_n - \widehat{\widetilde{\boldsymbol{\beta}}}_n) + o_P(1).\end{aligned}$$

Then the second part of Lemma 2 is proved. □

We now show Theorem 4. A direct use of Lemma 2 and (A.12) leads to

$$2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n)\} = \left(\frac{1}{\sqrt{n}}\mathbf{H}_n^{-1/2}\sum_{i=1}^n \mathrm{q}_{1,i}\widetilde{\mathrm{X}}_{ni}\right)^T H_n\left(\frac{1}{\sqrt{n}}\mathbf{H}_n^{-1/2}\sum_{i=1}^n \mathrm{q}_{1,i}\widetilde{\mathrm{X}}_{ni}\right) + o_P(1).$$

Since H_n is idempotent of rank k, it can be written as $H_n = C_n^TC_n$, where C_n is a $\mathrm{k}\times(p_n+1)$ matrix satisfying $C_nC_n^T = \mathbf{I}_{\mathrm{k}}$. Then

$$2n\{\ell_n(B_n^T\widehat{\boldsymbol{\gamma}}_n) - \ell_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n)\} = \left(\frac{1}{\sqrt{n}}C_n\mathbf{H}_n^{-1/2}\sum_{i=1}^n \mathrm{q}_{1,i}\widetilde{\mathrm{X}}_{ni}\right)^T\left(\frac{1}{\sqrt{n}}C_n\mathbf{H}_n^{-1/2}\sum_{i=1}^n \mathrm{q}_{1,i}\widetilde{\mathrm{X}}_{ni}\right) + o_P(1).$$

When the q-function satisfies (3.3), $\mathbf{H}_n = \Omega_n/c$. In this case, similar arguments for Theorem 2 yield

$$\frac{1}{\sqrt{n}}C_n\mathbf{H}_n^{-1/2}\sum_{i=1}^n \mathrm{q}_1(Y_{ni};\widetilde{\mathrm{X}}_{ni}^T\widetilde{\boldsymbol{\beta}}_{n;0})\widetilde{\mathrm{X}}_{ni} \xrightarrow{\mathcal{L}} N(\mathbf{0}, c\mathbf{I}_{\mathrm{k}}),$$

which completes the proof. □

Proof of Theorem 5

Note that W_n can be decomposed into 3 additive terms,

$$\begin{aligned}I_1 &= n\left\{A_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0})\right\}^T(A_n\widehat{V}_nA_n^T)^{-1}\left\{A_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0})\right\},\\ I_2 &= 2n(A_n\widetilde{\boldsymbol{\beta}}_{n;0})^T(A_n\widehat{V}_nA_n^T)^{-1}\left\{A_n(\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0})\right\},\\ I_3 &= n(A_n\widetilde{\boldsymbol{\beta}}_{n;0})^T(A_n\widehat{V}_nA_n^T)^{-1}(A_n\widetilde{\boldsymbol{\beta}}_{n;0}).\end{aligned}$$

We observe that $I_1 \xrightarrow{\mathcal{L}} \chi^2_k$ following the second part of Lemma 1; $I_3 = n(A_n\widetilde{\boldsymbol{\beta}}_{n;0})^T\mathbf{M}^{-1}(A_n\widetilde{\boldsymbol{\beta}}_{n;0})\{1+o_P(1)\}$ by Proposition 2; $I_2 = O_P(\sqrt{n})$ by Cauchy–Schwartz inequality. Thus

$$n^{-1}I_3 \geq \lambda_{\min}(\mathbf{M}^{-1})\|A_n\widetilde{\boldsymbol{\beta}}_{n;0}\|^2\{1+o_P(1)\} = \lambda^{-1}_{\max}(\mathbf{M})\|A_n\widetilde{\boldsymbol{\beta}}_{n;0}\|^2 + o_P(1).$$

These complete the proof for W_n. □

Proof of Theorem 6

Following the second part of Lemma 1, we observe that $\sqrt{n}(A_n\widehat{V}_nA_n^T)^{-1/2}(A_n\widehat{\widetilde{\boldsymbol{\beta}}}_n) \xrightarrow{\mathcal{L}} N(\mathbf{M}^{-1/2}\boldsymbol{c}, \mathbf{I}_k)$, which completes the proof. □

Proof of Theorem 7

We first need to show Lemma 3.

Lemma 3. *Suppose that* (X_n^o, Y_n^o) *follows the distribution of* (X_n, Y_n) *and is independent of the training set* $\mathcal{T}_n$. *If* $\mathtt{Q}$ *is a* BD, *then*

$$E\{\mathtt{Q}(Y_n^o, \widehat{m}(\mathrm{X}_n^o))\} = E\{\mathtt{Q}(Y_n^o, m(\mathrm{X}_n^o))\} + E\{\mathtt{Q}(m(\mathrm{X}_n^o), \widehat{m}(\mathrm{X}_n^o))\}.$$

Proof. Let $\mathtt{q}$ be the generating function of $\mathtt{Q}$. Then

$$\begin{aligned}\mathtt{Q}(Y_n^o, \widehat{m}(\mathrm{X}_n^o)) = {} & [\mathtt{q}(m(\mathrm{X}_n^o)) - E\{\mathtt{q}(Y_n^o) \mid \mathcal{T}_n, \mathrm{X}_n^o\}] + [E\{\mathtt{q}(Y_n^o) \mid \mathcal{T}_n, \mathrm{X}_n^o\} \\ & - \mathtt{q}(Y_n^o)] - \mathtt{q}(m(\mathrm{X}_n^o)) + \mathtt{q}(\widehat{m}(\mathrm{X}_n^o)) + \{Y_n^o - \widehat{m}(\mathrm{X}_n^o)\}\mathtt{q}'(\widehat{m}(\mathrm{X}_n^o)).\end{aligned} \tag{A.14}$$

Since (X_n^o, Y_n^o) is independent of $\mathcal{T}_n$, we deduce from [32, Corollary 3, p. 223] that

$$E\{\mathtt{q}(Y_n^o) \mid \mathcal{T}_n, \mathrm{X}_n^o\} = E\{\mathtt{q}(Y_n^o) \mid \mathrm{X}_n^o\}. \tag{A.15}$$

Similarly,

$$E\{Y_n^o\mathtt{q}'(\widehat{m}(\mathrm{X}_n^o)) \mid \mathcal{T}_n, \mathrm{X}_n^o\} = E(Y_n^o \mid \mathrm{X}_n^o)\mathtt{q}'(\widehat{m}(\mathrm{X}_n^o)) = m(\mathrm{X}_n^o)\mathtt{q}'(\widehat{m}(\mathrm{X}_n^o)). \tag{A.16}$$

Applying (A.15) and (A.16) to (A.14) results in

$$E\{\mathtt{Q}(Y_n^o, \widehat{m}(\mathrm{X}_n^o)) \mid \mathcal{T}_n, \mathrm{X}_n^o\} = E\{\mathtt{Q}(Y_n^o, m(\mathrm{X}_n^o)) \mid \mathrm{X}_n^o\} + \mathtt{Q}(m(\mathrm{X}_n^o), \widehat{m}(\mathrm{X}_n^o))$$

and thus the conclusion. □

Now show Theorem 7. Setting $\mathtt{Q}$ in Lemma 3 to be the misclassification loss gives

$$\begin{aligned}1/2[E\{R(\widehat{\phi}_n)\} - R(\phi_{n,\mathrm{B}})] \leq {} & E[|m(\mathrm{X}_n^o) - 0.5|\mathrm{I}\{m(\mathrm{X}_n^o) \leq 0.5,\ \widehat{m}(\mathrm{X}_n^o) > 0.5\}] \\ & + E[|m(\mathrm{X}_n^o) - 0.5|\mathrm{I}\{m(\mathrm{X}_n^o) > 0.5,\ \widehat{m}(\mathrm{X}_n^o) \leq 0.5\}] \\ = {} & I_1 + I_2.\end{aligned}$$

For any $\epsilon > 0$, it follows that

$$\begin{aligned}I_1 = {} & E[|m(\mathrm{X}_n^o) - 0.5|\mathrm{I}\{m(\mathrm{X}_n^o) < 0.5 - \epsilon,\ \widehat{m}(\mathrm{X}_n^o) > 0.5\}] \\ & + E[|m(\mathrm{X}_n^o) - 0.5|\mathrm{I}\{0.5 - \epsilon \leq m(\mathrm{X}_n^o) \leq 0.5,\ \widehat{m}(\mathrm{X}_n^o) > 0.5\}] \\ \leq {} & P\{|\widehat{m}(\mathrm{X}_n^o) - m(\mathrm{X}_n^o)| > \epsilon\} + \epsilon\end{aligned}$$

and similarly, $I_2 \leq \epsilon + P\{|\widehat{m}(\mathrm{X}_n^o) - m(\mathrm{X}_n^o)| \geq \epsilon\}$. Recall that

$$|\widehat{m}(\mathrm{X}_n^o) - m(\mathrm{X}_n^o)| = |F^{-1}(\widetilde{\mathrm{X}}_n^{o\,T}\widehat{\widetilde{\boldsymbol{\beta}}}_n) - F^{-1}(\widetilde{\mathrm{X}}_n^{o\,T}\widetilde{\boldsymbol{\beta}}_{n;0})| \leq |(F^{-1})'(\widetilde{\mathrm{X}}_n^{o\,T}\widetilde{\boldsymbol{\beta}}_n^*)|\|\mathrm{X}_n^o\|\,\|\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}\|,$$

for some $\widetilde{\boldsymbol{\beta}}_n^*$ between $\widetilde{\boldsymbol{\beta}}_{n;0}$ and $\widehat{\widetilde{\boldsymbol{\beta}}}_n$, where $\widetilde{\mathrm{X}}_n^o = (1, \mathrm{X}_n^{o\,T})^T$. By Condition A4, we conclude that $(F^{-1})'(\widetilde{\mathrm{X}}_n^{o\,T}\widetilde{\boldsymbol{\beta}}_n^*) = O_P(1)$. This along with $\|\widehat{\widetilde{\boldsymbol{\beta}}}_n - \widetilde{\boldsymbol{\beta}}_{n;0}\| = O_P(1)$ and $\|\widetilde{\mathrm{X}}_n^o\| = O_P(\sqrt{p_n})$ implies that $|\widehat{m}(\mathrm{X}_n^o) - m(\mathrm{X}_n^o)| = O_P(r_n\sqrt{p_n}) = o_P(1)$. Therefore $I_1 \to 0$ and $I_2 \to 0$, which completes the proof. □

References

[1] W.C. Hamilton, The revolution in crystallography, Science 169 (1970) 133–141.
[2] P.J. Huber, Robust regression: asymptotics, conjectures, and Monte Carlo, Ann. Statist. 1 (1973) 799–821.
[3] F.C. Drost, Generalized chi-square goodness-of-fit tests for location-scale models when the number of classes tends to infinity, Ann. Statist. 17 (1989) 1285–1300.
[4] S.A. Murphy, Testing for a time dependent coefficient in Cox's regression model, Scand. J. Statist. 20 (1993) 35–50.
[5] B.D. Ward, Deconvolution analysis of fMRI time series data, Technical Report, Biophysics Research Institute, Medical College of Wisconsin, 2001.
[6] K.J. Worsley, C.H. Liao, J. Aston, V. Petre, G.H. Duncan, F. Morales, A.C. Evans, A general statistical analysis for fMRI data, Neuroimage 15 (2002) 1–15.
[7] Y.H. Chen, S.L. Bressler, K.H. Knuth, W.A. Truccolo, M.Z. Ding, Stochastic modeling of neurobiological time series: power, coherence, Granger causality, and separation of evoked responses from ongoing activity, Chaos 16 (2006) Art. No. 026113.
[8] C.M. Zhang, T. Yu, Semiparametric detection of significant activation for brain fMRI, Ann. Statist. 36 (2008) 1693–1725.
[9] G.H. Glover, Deconvolution of impulse response in event-related BOLD fMRI, Neuroimage 9 (1999) 416–429.
[10] R.W. Cox, AFNI: software for analysis and visualization of functional magnetic resonance neuroimages, Comput. Biomed. Res. 29 (1996) 162–173.
[11] D. Cordes, R.R. Nandy, Estimation of the intrinsic dimensionality of fMRI data, Neuroimage 29 (2006) 145–154.
[12] J. Friedman, On bias, variance, 0/1-loss, and the curse-of-dimensionality, Data Min. Knowl. Discov. 1 (1997) 55–77.
[13] R. Tibshirani, Bias, variance and prediction error for classification rules, Technical report, Statistics Department, University of Toronto, 1996.
[14] G. James, Variance and bias for general loss functions, Mach. Learn. 51 (2003) 115–135.
[15] B. Efron, The estimation of prediction error: covariance penalties and cross-validation (with discussion), J. Amer. Statist. Assoc. 99 (2004) 619–642.
[16] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, J. Amer. Statist. Assoc. 101 (2006) 138–156.
[17] S. Portnoy, Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity, Ann. Statist. 16 (1988) 356–366.
[18] J. Fan, H. Peng, Nonconcave penalized likelihood with a diverging number of parameters, Ann. Statist. 32 (2004) 928–961.
[19] H. Cramér, Mathematical Methods of Statistics, Princeton University Press, Princeton, NJ, 1946.
[20] C.J. Geyer, On the asymptotics of constrained *M*-estimation, Ann. Statist. 22 (1994) 1993–2010.
[21] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.
[22] L.M. Brègman, A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (1967) 620–631.
[23] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, 2001.
[24] R.W.M. Wedderburn, Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method, Biometrika 61 (1974) 439–447.
[25] M.S. Bartlett, Approximate confidence intervals, Biometrika 40 (1953) 12–19.
[26] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.
[27] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Ann. Statist. 32 (2004) 56–134.
[28] P.L. Purdon, V. Solo, R.M. Weissko, E. Brown, Locally regularized spatiotemporal modeling and model comparison for functional MRI, Neuroimage 14 (2001) 912–923.
[29] C.M. Zhang, Y. Lu, T. Johnstone, T. Oakes, R.J. Davidson, Efficient modeling and inference for event-related fMRI data, Comput. Statist. Data Anal. 52 (2008) 4859–4871.
[30] D. Harrison, D.L. Rubinfeld, Hedonic housing prices and the demand for clean air, J. Environ. Econ. Manage. 5 (1978) 81–102.
[31] G.H. Golub, C.F. Van Loan, Matrix Computations, third ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[32] Y.S. Chow, H. Teicher, Probability Theory, second ed., Springer-Verlag, 1989.